

Progress on patents, but more action is needed

Europe's proposed directive on biotechnology patenting has passed its most difficult hurdle. But patent policy still faces problems, such as the way some claims are implemented, and the lack of a 'grace period'.

Opponents of biotechnology in Europe should soon have less opportunity to challenge it for bad reasons rather than good, while the clearing of an embarrassing backlog in biotechnology patent applications should soon be under way. That is the good news to come from last week's approval by the European Parliament of a draft directive on the legal protection of biotechnological inventions (see page 314).

The tightly worded directive makes it clear that, provided an inventive step and an appropriate use can be demonstrated, isolated human genes, as well as transgenic animals and plants, can be patented. Discoveries, such as a gene in its natural state, are thus not patentable. The directive also closes a technical loophole, relating to the need to protect plant breeders' interests, which opponents to patents on life had exploited to put a temporary stop to transgenic patents in Europe.

In the future, assuming that the European Patent Office and member states act — as expected — in accordance with the directive, critics will be restricted to using its 'morality clause' to oppose individual patents. That clause applies only to animals, not plants. It requires patent examiners to weigh up the level of suffering or physical handicap of a transgenic animal against its substantial benefit to human or veterinary medicine. Such qualitative criteria are, of course, open to interpretation, and will keep patent lawyers in pocket for years to come.

Despite this progress, researchers can expect continued opposition. Critics will keep up their fight outside patent office court rooms. They are likely to maintain political pressure to try to overturn or modify the directive — and will also seek other, more direct, ways to stop genetic engineering, the real motivation behind some opposition to biotechnology patents, at least those covering animals. A referendum planned

for next year may well lead to a complete ban on the use of transgenic organisms in Switzerland (see page 315). Some activists, even the most well informed, will no doubt continue to exploit public ignorance. One of the more vociferous of Germany's Green MEPs, for example, still argues that the directive will in practice allow patents to be issued on discoveries rather than inventions, despite assurances to the contrary.

What else remains to be done? The directive should not, and will not, put an end to ethical debate. But patent-holders must also learn to exploit their rights appropriately. Regrettably, the actions of at least one prominent company give cause for concern. Du Pont holds a patent on an inducible promoter, *Cre-loxP*, which has become a fundamental tool in the creation of transgenic mice. By adopting an excessively aggressive licensing policy, which in practice restricts the academic exchange of transgenic mice by placing a heavy administrative burden on researchers licensed to use them, Du Pont is making enemies unnecessarily. Indeed, scientists in many smaller laboratories are ignoring the licensing requirement, and seeking alternative methodologies.

There is another step that needs to be taken. European biotechnology remains disadvantaged by the absence of the so-called 'grace period' which US researchers enjoy between scientific publication and a deadline for a patent application. As a result, some European scientists feel compelled to delay publication of their genetic data for fear of losing patent rights. There is a strong argument for linking the introduction of a grace period in Europe (which would also benefit US inventors) to a similarly significant change in the US patenting system, awarding patent rights on the basis of 'first to file', rather than 'first to invent' as at present. Previous negotiations on this ran into the sand; they should be revived. □

Planting a sound idea

The US Senate should be applauded for rejecting the administration's caution on plant genome sequencing.

The launching of a plant genome initiative by Senator Christopher Bond (see page 312) is a good example of a vanishing phenomenon: a directive by the US Congress to get on with a project which the science agencies in the government know needs doing, but lack the resolve to proceed with.

Plant genome research, as the corn-growers who have badgered Bond into launching this initiative recognize, is a field of vast economic potential. Agricultural productivity in the United States has doubled in the past fifty years, and agriculture has become the unsung powerhouse of the American economy, generating an astounding \$30 billion trade surplus each year. Biotechnology is the most promising route to future gains.

Bond's proposal to kick-start the initiative with \$40 million in the next financial year from the budget of the National Science Foundation (NSF) is causing understandable concern in that agency. It will consume half of the NSF's annual budget increase, limiting other opportunities across its \$2 billion portfolio of university research.

But the senator has not had much choice. Ideally, the initiative should be paid for by the US Department of Agriculture (USDA), whose mission it will clearly serve. But the USDA's research operations are notoriously badly run. The Clinton administration, like its predecessors, has done little to fix them, and drew up some plans for a multi-agency plant genome effort only in April, when Bond asked it to do so.

The budget process now requires that Bond, his counterparts in the House of Representatives, and the administration should agree a structure for the initiative in time for the start of the next financial year in October. That should not be too difficult.

The NSF has, as its director points out, a duty to balance its portfolio carefully — but the portfolio is not set in stone. The new availability of sequencing technology cries out for a plant genome initiative at least as bold as Bond's proposal. NSF management will help to ensure the quality of the work, but the White House must ensure that, eventually, USDA bears a substantial share of the cost, and that the rest of science does not suffer. □



Pressure grows for inquiry into welfare of transgenic animals

[LONDON] Public concern in Britain over the increasing use of animals in genetics research is fuelling calls for a commission of inquiry to investigate the welfare of transgenic animals. The calls come in the wake of the UK government's decision last month to shelve an earlier promise to set up a broad-ranging royal commission on the use of animals in medical research.

Organizations on both sides of the animal experiments debate say they are relieved that plans for a wide-ranging royal commission have been abandoned. But many now want the government to focus on transgenic animals, an issue they believe is of more immediate public concern.

A recent survey supported by the European Commission into public attitudes to biotechnology showed that most respondents in European Union countries consider the creation of transgenic animals, for research and for xenotransplants, to be morally unacceptable (see *Nature* 387, 845–847; 1997).

The issue was highlighted in last week's debate in the European Parliament on biotechnology patent legislation, and has also become central to a national referendum on the applications of genetic engineering to be held in Switzerland next year (see pages 314 and 315). Abby Munson, a spokeswoman for the Genetics Forum, says governments can no longer ignore such concerns. "The science is progressing at such a speed that the public feels left behind."

Significantly, in the United Kingdom at least, while the numbers of animals involved in research has been dropping, the numbers of transgenic animals bred increased fourfold between 1990 and 1995, when 215,000 animals were produced. This figure is expected to increase further when statistics for 1996 are released by the government today (24 July).

The proposal for a commission on transgenic animals has come from Maggy Jennings, head of the 'animals in research' division of the Royal Society for the Prevention of Cruelty to Animals. The idea has many supporters, including the Fund for the Replacement of Animals in Medical Experiments (FRAME), the pressure groups, Genetics Forum and Advocates for Animals.

The idea of studying the implications of the increasing use of transgenic animals also has support from a number of scientists, including Colin Blakemore, professor of physiology at the University of Oxford.

Under the 1986 Animals (Scientific Procedures) Act — widely considered to be among the toughest pieces of legislation in the world, and which is currently being reviewed — all animals used in research need to be licensed by the Home Office's Animal Procedures Committee. The legislation is framed according to a cost-benefit approach — a procedure will be licensed if the benefits to humans, or other animal species, outweigh the costs, in terms of suffering, to the animal used in the research.



Pigs in the middle: Switzerland will vote next year on banning transgenic experiments (see page 315).

But Blakemore acknowledges that since the act was passed, in some instances the costs to the animals used have been increasing along with the "enormous" benefits to humans. "In some cases, the potential suffering may have increased to the point where it may be unacceptable. It is almost like saying 'would it ever be justifiable to kill people if the benefit was a guaranteed cure for cancer?'. The answer is, of course, 'no', but that is the direction we're heading in."

Blakemore says he would welcome an inquiry into transgenic animals, rather than a broad-ranging royal commission on animal testing, as the latter would reopen issues that have already been settled. But he adds that inquiries need to have a bottom line — namely that the use of animals in research should not be banned, as ultimately there will always be benefits to humans.

Some in the animal rights movement would still prefer a broad royal commission on animals in research. The British Union for the Abolition of Vivisection believes that neither an inquiry into transgenic animals nor a review of the animals procedures act is a proper substitute for a broad and objective inquiry into the whole question.

Mike Baker, the union's director, does not believe that the current review of the act will be sufficiently objective, as he says the government's Animal Procedures Committee — which is carrying out the review — is dominated by individuals with a research background, or at least an interest in research.

Although not opposed to an inquiry into transgenic animals, he argues that the issue was adequately dealt with by two reports on xenotransplants published earlier this year. One was commissioned by the Department of Health under the previous Conservative government, the other was

CNRS gets first woman director general

[PARIS] Catherine Bréchnignac, a physicist, last week became the first woman director general of the French Centre National de la Recherche Scientifique (CNRS), succeeding Guy Aubert, whose three-year term had expired.

The new Socialist government also appointed Gérard Brachet as director general of the French space agency, CNES, a post that had been abolished in January last year (see *Nature* 379, 476; 1996).

Bréchnignac, who is 51 years old, is well-respected as a scientist — she has worked on atomic clusters — and as a manager. Since November 1995 she has headed CNRS's Department of Physical Sciences and Mathematics.

Her task as director general is likely to be substantial, given that Claude Allègre, the dynamic minister for education, research and technology, is keen to see a broad reform

of research strategy within CNRS, as well as in universities and other public research organizations (see *Nature* 388, 7; 1997).

Allègre has complained that CNRS has become "an immense bureaucracy", and he intends to reduce the organization's central management and to transfer staff to reinforce the university system.

Allègre's unhappiness with France's space policy suggests that Brachet's job will be no less difficult. One task is expected to be to reorientate the space agency's activities so that they are driven less by politics and more by "technical and scientific aspects".

In a recent interview with the newspaper *Le Monde*, Allègre described the agency's policies as "stupefying", criticizing the excessive weight of its contribution to the European Space Agency and complaining that CNES's own objectives were far from clear.

Declan Butler

from the Nuffield Council on Bioethics.

Ian Kennedy, professor of law and medical ethics at King's College, London, who chaired the xenotransplants panel commissioned by the previous government, says the question of the ethics of transgenic animals was looked at in detail. But David Shapiro, executive secretary of the Nuffield Council on Bioethics, says that the issue of transgenic animals was only "touched on" in the council's report, and is oriented towards their use as a source of xenografts.

Baker argues that the topic of transgenic animals will resurface in a review of biotechnology and the patenting of animals which was also promised in the Labour party's pre-election policy document, *New Labour, New Life for Animals*. Indeed, some are pressing Britain's science minister, John Battle, to raise this issue when the European patent directive is discussed by the Council of Ministers.

But Baker's could be a minority voice. Les Ward, director of Advocates for Animals, says his experience of royal commissions is that "they tend to be talking-shops with little impact. Governments sometimes use them to deflect public opinion."

Some organizations say they were never enthusiastic about the proposal, but kept quiet about their reservations in an attempt not to embarrass Elliot Morley, the former Labour spokesman on animal welfare, whose idea it was. Morley is now a junior minister responsible for 'fish and countryside' in the Ministry of Agriculture, but is no longer responsible for animal welfare.

But the decision not to proceed with the royal commission remains controversial because it was understood to have been a pre-election promise. Some Labour members of parliament — including Anne Campbell, recently appointed as a parliamentary private secretary to Battle — appear to have been taken by surprise by the government's decision not to proceed.

Others are concerned that the decision could be interpreted as an indication that the new government makes promises in haste, and then fails to keep them. Lynne Jones (Labour, Birmingham Selly Oak) says she was "somewhat taken aback" by the decision, having thought that a commission was official Labour party policy.

Indeed, three weeks before the 1 May general election, Morley confirmed that Labour had promised a royal commission to look at the whole issue of animal experimentation.

Government sources now argue that the royal commission was never a firm manifesto promise, but was rather a promise made in *New Labour, New Life for Animals* to "support" a royal commission.

Civil servants are understood to have been concerned that taking on a royal commission would have created additional stress in what is already an ambitious legislative timetable.

Ehsan Masood

NSF urged to increase plant genome sequencing effort

[WASHINGTON] A \$40 million-a-year initiative at the US National Science Foundation to sequence a plant genome has been proposed by a key Senate committee, and could be under way by the end of this year.

The initiative is being vigorously promoted by Senator Christopher Bond (Republican, Missouri) at the prompting of US corn-growers, who are excited about the potential of genetically engineered crops to increase yields.

Scientists are convinced that the proposed initiative would support first-rate science and speed up the sequencing of the genome of corn and other crops. But some are also concerned that the proposal is so large that every other area of university science at the NSF would lose out to pay for it.

Last week, Bond earmarked \$40 million for a plant genome initiative in the budget bill prepared by the appropriations subcommittee of the Senate Veterans Affairs, Housing and Urban Development and independent agencies (VA-HUD), which he chairs.

Bond claims that the project is "key to keeping America number one in agriculture in the next century". It is said to be his top personal priority in the entire \$40-billion VA-HUD bill.

Agricultural interests, encouraged by the generally positive response to last summer's introduction of genetically engineered cotton crops (see *Nature* 387, 221; 1997), and aware of Japan's recent pledge to sequence the rice genome, are pushing for a coordinated effort corresponding to the Human Genome Project in the biomedical field.

The main concern of scientists is that Bond's initiative would consume half the extra money his committee wants to make available to the NSF next year, leaving other fields of science with an increase less than the rate of inflation, and dashing hopes that the science agency would have more money for general university research this year.

Together with the energy and agriculture departments, the NSF is engaged in a \$6-million-a-year project to sequence the genome of *Arabidopsis*, a mustard plant selected for early study because of its relatively small genome. Last year, this project was incorporated into a global effort.

Neal Lane, director of the NSF, pointed out last week that the agency already spends about \$20 million on plant genome research in total. But a member of Bond's staff says that the \$40 million is expected to pay for new work on top of the existing activity.

Officials from the NSF, the White House and both houses of Congress are expected to hold negotiations between now and Septem-



Genetic futures: companies like DeKalb Genetics Corp (above) may gain from sequencing project.

ber, when the budget is finalized, on the final shape of the initiative. "We all agree that this is something we should be doing," Bond's aide says. "It is not something we're going to fight about."

In April, Bond's subcommittee asked the White House to create an interagency working group on plant genomes. In a preliminary report last month, the group rejected an immediate project to sequence the corn genome on cost and benefit grounds. But it did endorse the principle of a plant genome initiative, to be led, it suggested, by the US Department of Agriculture, to lay the groundwork for future sequencing efforts.

Corn has a genome six times the size of that of rice and 17 times the size of that of *Arabidopsis*. The wheat genome is much larger still: six times larger than the corn genome. The working party said that the *Arabidopsis* work would help development of the powerful computer tools that would be needed to handle the larger genomes of the crop plants.

The NSF has traditionally fought off attempts by the Congress to tell it what research to do. But in this case, concern about the agency's independence may be alleviated by a widespread perception that plant genome research has been unfairly neglected in comparison to human genome research, which now has an institute of its own at the National Institutes of Health with an annual budget of almost \$200 million.

"We've been trying to make the argument that plant biotechnology is important for 20 years," says Mary Clutter, associate director of life sciences at the NSF.

Colin Macilwain

AP/BOB CHILD

Call for human subjects monitoring body

[SAN FRANCISCO] The US National Bioethics Advisory Committee is to examine the possibility of setting up a single, independent body to oversee the protection of human subjects in all federally funded research.

At a meeting of the committee's human subjects subcommittee in Bethesda, Maryland, last week, Alexander Capron, co-director of the Pacific Center for Health Policy and Ethics at the University of Southern California Law Center, and a former executive director of a presidential ethics committee under President Jimmy Carter, proposed commissioning papers to examine the matter.

Capron suggested the Office of Government Ethics as a potential model for an agency to oversee research. This became an independent body in 1989, its role being to maintain high ethical standards within executive branch agencies and departments.

The independent organization being considered by the bioethics committee would monitor problems in research on human subjects, and would be able to demand compliance by government bodies.

Capron argues that there has always been an inherent conflict within agencies that monitor their own research. As the rules that govern informed consent and other protective measures become more burdensome, officials at the National Institutes of Health have had to increase their emphasis on promoting the research. If research within an agency is overseen by that agency itself, he says, "I'm sceptical it will ever be possible to have as vigorous and thorough attention to

protection of the human side as to moving research forward".

An independent agency or department could protect human subjects without such pressure, Capron says. The committee has commissioned one paper to examine the benefits of such an approach and another to define the drawbacks. For instance, there is a risk that Congress would reduce an independent entity's funding or narrow its authority, or that the agency could become dominated by a particular area of research.

The need to get on with trials could also lose priority. Project Informed, an AIDS research project in San Francisco, has voiced concern that an independent body could get bogged down in bureaucracy.

At the recommendation of David Cox, professor of genetics and pediatrics at Stanford University and a member of the subcommittee, a third paper will consider whether private research should be in the new office's purview. Cox has argued that the division between private and public research has become artificial.

Cox warns that clear protective measures for human subjects need to be in place for the public to feel comfortable about taking part in research. In one study of genetic testing, for example, 30 per cent of patients refused to participate because they were worried about privacy.

Capron notes that a bill introduced by Senator John Glenn (Democrat, Ohio) would require private research laboratories to register with the Health and Human Ser-

vices Department. Glenn told Congress he was concerned about four major gaps in the current system to protect human subjects.

The four gaps were: that not all agencies, including the Department of Labor and the Nuclear Regulatory Commission, had adopted the common rule for human subjects research; that federally funded institutions needed to apply the policy only to funded research; that private institutions did not need to comply; and that only Health and Human Services had special protective measures for vulnerable categories of people such as children and prisoners.

Glenn pointed to examples reported in the *Cleveland Plain Dealer* and elsewhere in the media where people did not know they were taking part in research or did not know about potential side-effects.

According to Capron, private institutions may have their own internal office for protection of human subjects, but overall monitoring might be accomplished more effectively from the outside.

Gary Ellis, director of the Office of Protection from Research Risks at the National Institutes of Health, wants to extend legal protection to those taking part in private research projects that are not participating in the current voluntary system. **Sally Lehman**



John Glenn: wants labs to register.

Collins' student sanctioned over 'most severe' case of fraud

[SAN DIEGO] A former star student of the director of the Center for Human Genome Research at the US National Institutes of Health (NIH) has been barred from federally funded research for four years and denied his doctorate after one of the United States' biggest cases of scientific misconduct.

Amitov Hajra, a doctoral candidate who worked with the NIH genome chief Francis S. Collins in both Ann Arbor, Michigan, and Bethesda, Maryland, agreed to the penalty after an investigation by the NIH's Office of Research Integrity (ORI). Hajra was also studying for a medical degree at the University of Michigan.

Probes by Michigan and ORI established that Hajra had fabricated 75 to 90 per cent of the data in his doctoral thesis on a possible genetic cause of a leukaemia.

He then used fabricated data to produce five articles published in 1995 and 1996, included falsified data in two review articles, and entered a bogus nucleotide sequence in GenBank, a computer database. All the fake

research has been retracted or corrected. Neither Collins nor other co-authors were found to have had any involvement in the misconduct (see *Nature* 384, 6; 1996).

In their report delivered in March, Michigan officials referred to the "severe impact" the case had on "the public's confidence in science". They wrote: "On a scale of 1 to 10, with 10 being the most severe, the committee concludes that the academic misconduct here rates 10."

Hajra, who is now in Wisconsin, could not be reached for comment.

Michigan officials found the fabrications began when Hajra "suddenly blurted out" fake experimental results at a laboratory meeting. Afterwards, the report says, "he states he was so ashamed and confused by his behavior that he was unable to consider confessing to his wrongdoing."

Collins, who had brought Hajra with him when he was recruited in 1993 to the NIH, says: "He was very motivated, determined and clever. I still don't know

how he produced some of the primary data he brought to lab meetings. I find myself now being more vigilant, viewing data with more scepticism, which I'm not sure is a good thing."

Chris B. Pascal, acting director of ORI, commended Collins for the forthright way he handled the misconduct, which Collins first disclosed in a widely distributed letter after a reviewer found possible fabrication in a Hajra paper submitted to the journal *Oncogene*. The paper was withdrawn.

"He [Collins] reacted as best he could, confronted the individual and got an admission," says Pascal. "His reaction was more positive than we get from a lot of big shots, who engage in denial and in some cases cover-up."

Michigan will not grant Hajra his doctorate in human genetics. But Hajra continues to fight to secure his medical degree. A Michigan spokeswoman said the university had not decided whether to award Hajra this qualification. **Rex Dalton**

Euro-vote lifts block on biotech patents

[MUNICH] A long-standing question at the European Patent Office (EPO) about whether it is legally entitled to grant patents on transgenic animals and plants could soon be resolved, following last week's approval by the European Parliament of revised rules on the legal protection of biotechnological inventions.

The EPO has almost 2,000 such applications in abeyance, following a decision two years ago that the exclusion of plant and animal varieties from patentability by the European Patent Convention of 1973 is sufficient to reject a patent on a plant or animal that has been genetically modified.

But the draft European directive explicitly allows patenting of life, including human genes, provided that a truly inventive step and industrial use can be proved, and provided that the procedures involved are within defined ethical limits.

It also requires the European Commission to set up a bioethics committee, which would report annually to the parliament on issues concerning the patenting of biotechnological inventions (see panel).

The directive must be approved by the European Council of Ministers in November, before returning to parliament for its second reading, expected early next year. But the council's support for the initial draft, the modifications to which have been generally accepted by the commission, means that a successful conclusion seems virtually guaranteed.

The pharmaceutical industry has welcomed the parliamentary vote, saying that it

will help Europe's biotechnology sector to catch up with the United States and Japan. But it has been criticized by environmentalists, who continue to oppose patents on novel life-forms.

The first draft of the directive was unexpectedly rejected by the parliament in 1995 because of ethical concern about patenting of life and human genes, as well as concern that it could impinge on farmers' privilege to breed or propagate stock (see *Nature* 374, 103; 1995).

The commission, which drafts all European Union directives, was more careful with the text in its second attempt. It consulted industry, pressure groups and other stakeholders before putting pen to paper (see *Nature* 379, 197; 1996), and included additional clauses to safeguard ethical standards.

But members of the parliament still tabled hundreds of amendments, reflecting a controversy that cuts across political party lines. And, in a unique move, the parliamentary Committee on Research, Technological Development and Energy declined to give an opinion of the directive before its first reading because of disagreements about which amendments to recommend.

As now approved by the parliament, the draft directive excludes patents on procedures for human reproductive cloning and germline therapy, and methods using human embryos. But many opponents of the directive remain upset that patenting of any form of life should be allowed. "This is a sad day for human respect for life," said a spokesman for Greenpeace after the vote.



Peter Stevenson, political and legal director of the British pressure group Compassion in World Farming (CIWF), says he is "deeply disappointed" that parliament has voted to allow animals to be patented. But he welcomes a clause that excludes patents on processes for modifying the genetic identity of animals "which are likely to cause them suffering or physical handicaps without any substantial medical benefit to man or animal".

The addition of the word 'medical' by the parliament means that benefits to agriculture alone will not necessarily outweigh any proven suffering to an animal genetically engineered, for example, to overproduce growth hormone, as grounds for rejecting a patent.

Stevenson says that CIWF has filed an opposition to a patent issued to the Australian company Bresagen on a method for producing transgenic pigs with extra growth hormone genes, claiming that such animals suffer joint disease, gastric ulcers and diabetes.

A spokesman for the EPO says the exclusion clause is likely to cause problems for patent lawyers because of the difficulty of legally interpreting what constitutes a physical handicap. In principle, the EPO is not formally bound by EU directives, as it is not an EU body. It was established by a separate convention with its own rules on what may be considered patentable. But in practice the directive would have a profound effect on EPO policy.

Eighteen months ago, opponents of patents on life, who had never managed successfully to oppose a patent on ethical grounds, exploited a loophole in the EPO's convention and put an abrupt stop to issuing of any new patents on transgenic plants and animals.

The convention, which was drawn up in the early 1970s, allows animals and plants to be patented in principle, but excludes animal and plant varieties from patentability, to allow breeders to claim their traditional non-patent rights to new varieties.

...but parliament wants closer scrutiny

[MUNICH] The European Parliament wants to have more influence over the debate on bioethics in Europe, and would like ethical committees to be more concerned about the effects of biotechnological research on society.

An amendment to the European Commission's draft directive on the legal protection of biotechnological inventions, which was approved last week at its first parliamentary reading (see above), urges the commission to establish a committee on the ethics of biotechnology. This would report to parliament annually on all ethical questions

relating to biotechnology. Details of its responsibilities have yet to be defined.

The proposal is closely connected to a recent parliamentary resolution extending to the end of the year the mandate of the commission's existing Group of Advisors on the Ethical Implications of Biotechnology (GAEIB), which was due to expire later this month.

The resolution called for the commission to renew and clarify GAEIB's role, in close consultation with parliament. It formally commended the group, which was set up in 1991, for contributing to the public debate on bioethics through its series of reports. But it

also argued that "too much attention has been paid to the interests of research and not enough to the possible effects on society".

The commission is likely to recommend to parliament that the committee demanded by the draft directive should be closely connected with, or identical to, GAEIB, but details have yet to be formulated.

Noëlle Lenoir, a member of the French constitutional court who chairs GAEIB, says that the group would welcome what she sees as a natural expansion of its role to cover ethical issues relating to patenting of biotechnological inventions.

A.A.

In 1995 an EPO board of appeals ruled against a patent issued to the Belgian company Plant Genetics Systems (PGS) on a procedure for producing plants containing a new gene conferring resistance to herbicides based on glutamine synthase inhibitors, which also covered their progeny. The patent office accepted opponents' arguments that a transgenic plant can also be considered a plant variety, and is therefore not patentable.

Until this precedent is overturned by a further decision by an appeals board, the EPO cannot legally issue further patents on transgenics. A backlog of around 1,200 patent applications for plants, and 500 or 600 on animals, has since built up.

The next case to be brought to the EPO board of appeals, scheduled for October, will consider an appeal from the Swiss company Novartis against the rejection — following the PGS case — of its patent application for a process for creating transgenic plants containing genes conferring pathogen resistance, and again covering their progeny.

The board will refer to the draft directive to help it to interpret the ambiguity in the EPO's rules created by the advent of genetic engineering. The draft states unambiguously that "inventions which concern plants or animals may be patented if the practicability of the invention is not technically confined to a particular plant or animal variety".

In the unlikely event that the board of appeals rejects this interpretation and upholds the ruling in the PGS case — that animals and plants are not legally distinguishable from their varieties — it would cause a major conflict, as EU member states, which form the core of EPO members, will be required to incorporate the directive in national legislation.

Alison Abbott

Swiss researchers facing 'anti-transgenics' vote

[MUNICH] Swiss scientists are warning that biomedical research in universities will be seriously harmed if a referendum next year calling for restrictions on genetic engineering — including a ban on the use of transgenic animals — wins approval.

A recent survey by the Swiss Societies for Experimental Biology (SSEB) found that at least 2,000 researchers involved in about 500 research projects would be directly affected by a proposed ban on the use of transgenic animals and field trials of genetically modified plants.

The referendum also proposes a ban on patenting genetically modified animals and plants, and demands that scientists demonstrate in advance that their research is of value to society and will not pose an environmental risk. Scientists would have to show that there are no alternative methods, and agree to observe approved ethical standards.

The survey indicates that Switzerland would quickly lose its high ranking in international genetics research if the proposals go through. "A ban on transgenic animals would be a virtual death blow for our research," says Rolf Zinkernagel, director of the Institute for Experimental Immunology at the University of Zurich, where 80 per cent of research projects involve transgenic animals.

Zinkernagel says that he would consider leaving Switzerland if the referendum resulted in a ban. "But the real victims would be the young generation of Swiss scientists who would be deprived of access to essential mol-

ecular biological techniques," he says.

The referendum was initiated in 1992 by 70 pressure groups opposed to biotechnology. According to Switzerland's direct democracy rules, any proposal for a constitutional amendment that secures the signatures of 1.5 per cent of the population must be put to a vote. If there is a simple majority, with a majority of cantons voting in favour, the proposed changes are incorporated into law.

The pharmaceutical industry has already launched an advertising campaign supporting the use of genetic engineering techniques in research in Switzerland. Christian Manzoni, a spokesman for a campaign launched by Hoffman-La Roche and Novartis, says that a 'yes' vote would be "a catastrophe for the industry". Manzoni warns that, if all research using genetic engineering techniques had to be done outside Switzerland, the companies would lose control over both ethical standards and methodologies.

University research departments are even more worried. "The pharmaceutical industry can take its genetic research abroad," says Peter Mani, director of the department of gene technology and society at SSEB. "But basic research at universities would suffer considerably from the proposed ban."

Previous surveys have shown that three-quarters of the Swiss population are against field trials of genetically modified organisms, novel foods and cloning, although a majority are in favour of medical applications of biotechnology.

Quirin Schiermeier

Russian minister tells academy: streamlining must continue

[MOSCOW] The Russian Academy of Sciences (RAS) is to merge some institutes whose directors are not RAS members with those headed by academicians. This follows a meeting with Vladimir Bulgak, vice prime minister for science and high technology in the Russian cabinet, who insisted that the number of inefficient institutes should be reduced so that the remaining institutes can be more effectively supported.

Bulgak appears determined to continue the reforms started by the former minister of science and technology policy, Boris Saltykov. Reforms had, however, stalled under Saltykov's successor, Vladimir Fortov. Fortov is vice-president of the RAS, academic secretary of its physics department and director of an academy institute. Many feel there is a conflict of interest between his academy role and his responsibility for the whole of the science budget.

In the new move, the Institute of

Spectroscopy in Troitsk, near Moscow, for example, which is relatively successful but whose director, Vladilen Letokhov, has failed to be elected to the academy, will be incorporated into the department of optics of the Physical Institute, which is in financial trouble but is headed by an academician.

The situation is further complicated by the recent election to the academy of Mikhail Kirpichnikov, chief of the cabinet's science and education department (see *Nature* 387, 535; 1997). Universities in particular are worried that Kirpichnikov will support the academy's interests but not necessarily those of other scientific institutions.

Bulgak told a recent press conference that the state will provide 13,000 billion rubles (US\$2,250 billion) a year for research, while the deputy finance minister, Izosim Molchanov, promised at a business meeting held by Bulgak that current debts to scientific institutions will be paid.

Bulgak is keen to convince others that change is necessary. The average salary of a Russian scientist is about 700,000 rubles (US\$120) a month. Trade unions want this increased to 1 million rubles. But Bulgak said that, to achieve this, "scientific institutions have only two options: earn more in Russia and abroad, or reduce staff".

His suggestion that some institutes should be closed has political opponents. According to Vladimir Babkin, an expert in science and education with the State Duma (the Russian parliament's lower chamber), "the government's attitude towards national science can be compared to that of the physician who said that the best remedy against headache is a guillotine".

But politicians have not suggested any other way out of the problems facing Russian science, and their criticism of the government is widely seen merely as a way of winning public sympathy.

Carl Levitin

Tokyo officials are told to agree on carbon strategy

[TOKYO] Japan's Prime Minister, Ryutaro Hashimoto, has called on senior ministers to sort out the country's policy on greenhouse gas emissions within two weeks, after growing international criticism of Japan's failure to make its position clear.

The vagueness of Japanese policy on reducing carbon dioxide emissions was criticized by Helmut Kohl, the German chancellor, at the recent meeting of the Group of Eight (G8) industrialized nations in Denver, Colorado. And there is increasing domestic concern about prospects for the United Nations climate change conference, to be held in Kyoto in December.



Ishii: needs a strategy on CO₂ emissions.

Last week, Hashimoto called on Michiko Ishii, head of the Environment Agency, Yukihiko Ikeda, the foreign minister, and Shizuka Kamei, acting minister of trade and industry, to come up with a strategy that will lead to agreement at the conference, at which more than 150 countries have already agreed to set legally binding emission reduction targets.

Ministers are expected to announce Japan's plans to deal with global warming at a cabinet meeting within the next few days. There have been widespread reports of disagreement between the pro-business Ministry of Trade and Industry and the Environment Agency about reductions of greenhouse gas emissions (*Nature* 387, 641; 1997).

Hironori Hamanaka, director general of the global environment department at the Environment Agency, denies there is a significant difference of opinion in the government. But he does not think Japan will have agreed reduction targets in time for an international preparatory meeting in Berlin next week.

Environment Agency officials also deny reports that Japan, together with Australia and the United States, is trying to persuade the European Union to withdraw its proposal for a reduction of 15 per cent in carbon dioxide emissions. Hamanaka says it is important for Japan to propose its own targets, but that a 15 per cent reduction is unrealistic.

Analysts in Tokyo think that without a revision to Japan's energy policy, it will be very difficult for the country to reduce carbon dioxide emissions at all. But international pressure on Japan is expected to mount in August when Germany's minister for the environment, Angela Merkel, visits Japan, and to intensify as the conference approaches.

Richard Nathan

Australia to rejoin space race with satellite project

[SYDNEY] Australia is making a belated attempt to join the competition for a new generation of microsatellites with the approval of government support for the creation of a cooperative research centre (CRC) for satellite systems.

The centre plans to build a 50-kg satellite for low-orbit experiments in communications, space physics, remote sensing and engineering. It will be launched in 2001 as an official part of the centenary celebrations of the federation of Australia's six states as an independent country, hence the project's name, FedSat-1.

With its base for military and civil tests at Woomera in the sparsely populated centre of the continent, Australia was in the forefront of space developments in the 1960s, a position enhanced by delays to the US manned flight programme after the fatal ground fire on Apollo 1.

In 1967, Australia became only the third country to launch a satellite from its own soil, placing one of the first X-ray telescopes in orbit. Built by the Weapons Research Establishment and the University of Adelaide, this was carried on an American military Redstone rocket.

But the rocket range fell into disuse after the cancellation of the British government's Blue Streak missile programme, which had been under test at Woomera, and the collapse of subsequent efforts to develop a civilian launcher through the European Launcher Development Organisation.

From the late 1980s to early 1996, the then

Labor government promoted Australia's potential as a launch site, whether for polar orbiting satellites from Woomera or near-equatorial launches from Cape York in the northeast. The bodies responsible for these efforts were disbanded to save money in the budget of August 1996.

But the science minister, Peter McGauran, retained A\$2 million (US\$1.5 million) a year and decided that a new CRC would specialize in space-related research. It was to be based on a group in the Commonwealth Scientific and Industrial Organization, the government's largest research agency, which is already carrying out remote sensing of Australia's vast land and ocean territory.

After delays in obtaining promises of collaboration from universities and small companies that produce components for satellites launched by other countries, the centre has now been approved, and its public funding increased to A\$20 million over seven years, which is more than matched by contributions of A\$36 million from the partners.

Design and construction of the microsatellite will take place mainly in Adelaide, Canberra and Brisbane. Although the launch site and vehicle have yet to be decided, the centre's director, Brian Embleton, says a firm offer of a rocket has been received from the Japanese Space Agency, with expressions of interest from Russia and the United States. Embleton says Asian countries are keen to share payloads with the Australian project as an affordable means of obtaining information from space.

Peter Pockley

NIH and genome project set for more funds

[WASHINGTON] The US National Institutes of Health (NIH) are on course for another substantial funding increase next year. Appropriators in the House of Representatives last week agreed a bill that will increase its budget by 6 per cent to \$13.5 billion, \$400 million more than requested by President Bill Clinton.

The increase, identified as a priority by both the chairman and the senior minority member of the Labor, Health and Human Services appropriations subcommittee, pleased lobbyists for biomedical research, who noted that the bill was passed without any unwanted earmarks or restrictive legislative riders.

The lobbyists are confident that the Senate appropriations subcommittee will approve an even larger increase for NIH when it addresses its own bill this week.

Senator Arlen Specter (Republican, Pennsylvania), who chairs the subcommittee, has said that he would try to secure a 7.5 per cent increase for the NIH.

The House bill proposes especially large increases for the National Institute of Human Genome Research — up by 12 per cent to \$211 million — and for the National Institute of Diabetes and Digestive and Kidney Diseases, up 7.5 per cent to \$874 million. These increases reflect Congress's enthusiasm for the Human Genome Project and its recent interest in diabetes research.

The bill passed by the House subcommittee also has modest increases for the Agency for Health Care Policy and Research, an agency targeted for deep cuts by conservatives in the previous Congress, and for the Centers for Disease Control and Prevention.

Colin Macilwain

Talks start on policing bio-weapons ban...

[PARIS] Prospects for adding a verification regime to the worldwide ban on the production, possession and use of biological weapons edged a step closer last week. Talks opened in Geneva on a 'rolling text' that will form the basis of negotiations.

But industries using biotechnology and genetic engineering techniques, which are coming under increasingly close scrutiny from arms control agencies, are concerned that compliance could lead to extra regulations and loss of confidential information.

Awareness has grown recently of the need to reinforce the 1975 United Nations Biological Weapons Convention (BWC) with measures to deter and detect cheats. For example, although the former Soviet Union was one of the sponsors of the convention, it has admitted running a covert weapons programme until 1992, while the Gulf War revealed that Iraq possessed biological weapons.

The lack of a verification protocol in the convention "is one of the main weaknesses in the international regime against weapons of mass destruction," says Graham Pearson, former head of the UK Chemical and Biological Defence Establishment at Porton Down. Pearson points out that both the Chemical Weapons Convention agreed earlier this year and the Nuclear Non-Proliferation Treaty have strict verification regimes.

The preliminary rolling text was prepared by four working groups for an *ad hoc* committee of around 50 countries chaired by Tibor Tóth, Hungarian ambassador to the United Nations, and remains confidential. But it is believed to provide only a skeleton for the protocol, with most of the text yet to be agreed. Flesh will be added during the current talks which end on 1 August, and in a further round in September, before negotiations on the details begin in earnest next year. A conference of the BWC countries is expected to be held in 1999 to adopt the protocol.

The text's provisions include mandatory declaration of all sites of most relevance to the convention, and the means to launch investigations of suspected facilities and incidents. These measures would be enforced by a permanent international secretariat. Provisions would apply to all relevant civil and military research and production centres, including those of the pharmaceutical and biotechnology industries.

Few believe that biotechnology companies in democratic countries would risk working on weapons, but submitting them to strict controls is seen as the price of a protocol that will allow checks on military installations and rogue countries. Site inspections and removal of samples could threaten commercial confidentiality, however, warns EuropaBio, a trade organization representing 600-plus European companies.

Andrew Dickson, its secretary general, says that the industry nonetheless "recognizes the need for a verification regime," and will work to find amicable solutions.

The Association of Pharmaceutical Research and Manufacturers of America (PhRMA) says it wants to work "to reduce the threat of biological warfare". But it will "oppose any protocol that does not fully protect confidential business information".

PhRMA says it is opposed to routine inspections, arguing that inspections should be permitted only in response to specific allegations of violations of the BWC. It says inspections should be authorized only after a majority vote by an executive council of government representatives.

The task of overcoming industrial concerns should be eased by the experience gained during negotiation of the Chemical

Weapons Convention (CWC), says Martin Kaplan, director of the Geneva office of Pugwash. One compromise in the CWC was the concept of "managed access", which allowed an inspected state to propose alternative means of demonstrating compliance, for example by offering spot checks on some but not all parts of an installation, but with the choice of which remaining with the inspectors. Another idea being floated is to forgo removal of samples during inspections, with all analyses being carried out on-site.

Kaplan says a major stumbling-block in the negotiations will be the financing of the verification regime, which could cost US\$100 million annually. Developing countries are also expected to hold agreement of the protocol hostage to obtaining technology transfer agreements on biotechnology from industrialized nations. **Declan Butler**

...as 'designer weapons' threat is disputed

[PARIS] A major spur to efforts to strengthen the Biological Weapons Convention is mounting concern about the spectre of 'designer' biological weapons. Experts are divided over the extent to which such weapons pose a realistic threat. But there is widespread consensus that the convention and verification measures need to discourage any attempts to develop them.

There are two broad classes of designer weapons: biotechnological weapons, where an organism's characteristics are modified to enhance its military potential, and genetic weapons that could target genetic differences between populations.

Concern about such weapons crystallized at the convention's fourth review conference in Geneva last winter. Although such weapons are prohibited under the general wording of the convention, the conference felt it necessary to extend this explicitly to include applications arising from "molecular biology" and "genome studies".

Concern about designer weapons has also recently been expressed by the

International Committee of the Red Cross, the Federation of American Scientists and the World Medical Association.

Vivienne Nathanson, head of science and ethics at the British Medical Association, says that, as genome research is aimed at finding targeting systems for drugs, such knowledge could also be used to inflict damage. "The potential for weaponization of that knowledge is significant," she says.

Weapons based on genetic manipulation – for example to enhance toxin production – are considered by many to be a realistic, if remote, threat. "It's not difficult to imagine pathogenicity being engineered into an organism, while designing an antidote specifically for your own troops," says Jean Pascal Zanders of the Stockholm International Peace Research Institute.

But genetic weapons are dismissed by many weapons experts as science fiction. The technical feasibility of developing them "cannot be discounted, but at this stage it is bunk," says Matthew Meselson of the Department

of Molecular and Cellular Biology at Harvard University, an expert on chemical and biological weapons.

Meselson is also sceptical that militarily useful genetic differences could be identified, pointing out that ethnic differences are often based on culture, not on genetics. At worst such weapons would "preferentially harm one group over another," he says.

Would-be proliferators are also likely to be discouraged from using designer weapons because they would need to be extensively tested in the field, says Graham Pearson, former head of the UK Chemical and Biological Defence Establishment.

The risk of designer weapons should nonetheless "not be downplayed", warns Barbara Hatch Rosenberg, a microbiologist at the State University of New York at Purchase and chair of the Federation of American Scientists' Chemical and Biological Weapons Program. "I think the risk should be taken seriously," she says, adding that the convention should take such risks into account. **D.B.**

International initiative completes sequence of *Bacillus subtilis*

[LONDON] The sequencing of 4.2 million bases of the bacterium *Bacillus subtilis*, an important source of industrial enzymes, and used for purposes ranging from the study of anthrax to the production of the traditional Japanese dish Natto, has been completed by an international consortium of laboratories. The work was funded by the European Commission in Brussels.

The participants, which include 28 labs from six European countries, have set up an 'industrial platform' with European biotechnology companies to explore the commercial applications of the genes discovered. The genome is known to contain several genes involved in the production of proteins with antibiotic properties.

The project, which began in September 1990, was coordinated at the Institut Pasteur in Paris and the Nara Institute of Life Sciences in Japan. Participants also included seven Japanese laboratories, two in the United States, and one in Korea. A new programme, also supported by the commission, has now been launched to define the function of the bacterium's 4,000 genes. Part of the data can be found at <http://www.pasteur.fr/Bio/SubtiList.html>.

More grants for Mars meteorite studies

[WASHINGTON] The US National Science Foundation last week announced the award of seven grants, worth a total of \$800,000, to university research groups to carry out further investigations into the claim, published last summer, that the martian meteorite ALH84001 once harboured life.

The awards, which come on top of 16 others made by the National Aeronautics and Space Administration last month, will enable university scientists to examine various aspects of the claim in closer detail. Some of the investigators will study small samples from the meteorite, while others will look for analogous features on terrestrial rock which may have borne life in extreme habitats.

US graduate students to keep tax break

[WASHINGTON] A threat to make research and teaching assistants at US universities pay thousands of dollars of tax on the free tuition they receive was receding this week, as negotiators in the House of Representatives and the Senate sought to agree a final version of major tax legislation. The House version of the tax proposal would have removed a 'tuition waiver' for these graduate students, requiring them to pay income tax on the

perquisite of free tuition — nominally worth \$20,000 or more a year — from their stipends of about \$10,000 a year.

Universities, which feared that they would have to raise stipends sharply to enable students to afford the new tax, joined with thousands of students in protesting against the tax change. According to several sources, including Richard Neal (Democrat, Massachusetts), the House side agreed last week to withdraw the change. But the bill may still scrap a similar waiver allowing university staff and their relations to pay no tax on their free tuition.

Brazilian soccer star gives kick for leprosy

[PARIS] The famous Brazilian footballer, Pelé, is to act as a 'goodwill ambassador' for the World Health Organization (WHO) to help publicize the fight against leprosy. Pelé — Edson Arantes do Nascimento — who scored 1,281 goals in 1,363 professional matches during his football career and is minister for sport in Brazil, already has a high profile in supporting leprosy control in Brazil. The country has 106,000 cases, second only to India. WHO has set the target for its Leprosy Elimination Campaigns to reduce prevalence to less than one case per 10,000 of the population, mainly by detecting undeclared cases and by providing multidrug therapy.

South Korea to back life sciences network

[TOKYO] The South Korean government is planning to call on the Asia-Pacific Economic Cooperation forum to support an initiative by life scientists to create a network in the region to promote collaboration in biotechnology and biomedical research.

Proposals for the network — the International Molecular Biology Network for Asia and the Pacific Rim — were discussed at a meeting in Tokyo in June (see *Nature* 388, 3; 1997). A representative of South Korea's Ministry of Science and Technology plans to express his government's intention to provide "significant levels of financial and technical support" for the network at a meeting of the forum's working group on industrial science and technology in Singapore in October, and to call on other countries to join it in supporting the network.

German space strategy stresses industry links

[MUNICH] The German cabinet last week approved a new space strategy for Germany. It emphasized closer collaboration with industry to promote the commercial exploitation of space, and tighter links with space agencies elsewhere, particularly the

European Space Agency (ESA).

The cabinet also promised to maintain the commitment made by Germany at ESA's ministerial meeting in Toulouse two years ago to become the lead player in the agency's contribution to the International Space Station. But the outline strategy itself was not accompanied by any specific details or indications of funding levels.

Japanese funds sought for new drilling ship

[TOKYO] A Japanese proposal to construct an ocean drilling ship to drill in more difficult geological environments, and to far greater depths, than the drill ship *JOIDES Resolution* of the current international Ocean Drilling Program (ODP), was discussed by Earth scientists from around the world at a meeting in Tokyo this week.

The proposed drill ship is the key component of a Japanese initiative called 'Ocean Drilling in the 21st Century' that is intended to be part of a future 'integrated ODP' once the current programme ends in 2003. Japan's Science and Technology Agency is planning to include the initiative in its budget proposal to be submitted at the end of August. The Tokyo meeting will discuss the project's scientific objectives and the strategies and technology needed. But it remains uncertain whether the proposal will

receive the support of Japan's Ministry of Finance in the government's current mood of fiscal restraint.

Pioneer 'astrogeologist' dies in Australian accident

[WASHINGTON] Eugene Shoemaker, one of the pioneers of planetary science, was killed in a car accident last week when his four-wheel-drive vehicle collided head-on with another vehicle on a rural road northwest of Alice Springs, Australia. His wife and colleague Carolyn Shoemaker was also injured, but was reported to be in a stable condition. The Shoemakers spent two weeks in the



Shoemaker: expert on planetary impacts

Australian outback every year investigating impact craters.

Eugene Shoemaker, who was 69, created the astrogeology branch of the US Geological Survey, in Flagstaff, Arizona, in 1961, and was instrumental in training the Apollo astronauts for geological fieldwork on the Moon. A recipient of the US National Medal of Science and numerous other awards, he was perhaps most influential in pointing out the significance — and threats — of planetary impacts.

The serious business of listing authors

Sir — The recent leading article on authorship (*Nature* 387, 831; 1997) struck a resonance within our team, as we have been wrestling with this demon for many years. Ours is an interdisciplinary undertaking, the success of which required extensive (and sometimes tedious) apparatus and software development before the first science results appeared (*Nature* 365, 621; 1993).

important contribution: the construction of the apparatus, the Terabyte database software, the time-series analysis, the detection efficiency determination, the comparison with models or the actual drafting of the paper? How should we compensate for the opportunity cost incurred by talented scientists who took on essential but time-consuming code development, forgoing the chance to publish 'idea' papers?

After considerable debate, we decided early on to use alphabetical authorship, because it may be awful but it's better than the alternatives (most of which seem to involve hand-to-hand combat between alleged colleagues). We have not been able to devise a scheme that is 'fairer', despite considerable creative effort.

Having made the decision to go with the alphabetical approach, we nevertheless seem condemned to revisit the topic every other year. Legitimate issues such as career advancement for junior team members, recognition for individual effort and initiative, and significant cultural differences between our disciplines remain sources of debate.

I am convinced that, in large collaborative undertakings that by their very nature make it difficult to decide the relative priority of individual contributions, the science is best served with alphabetical authorship. It is then incumbent on the

Which was the more

more senior members of the team to spell out the individual

contributions of students, postdocs and junior faculty members when recommendations are solicited, and to take active steps to draw attention to their efforts.

Concrete suggestions that might ease some of the problems include journal policies that would accommodate or even encourage papers authored by a team, with the individual aspects being secondary and less emphasis on first authorship for career advancement in disciplines where papers with 10 or more authors are not yet common.

Christopher Stubbs

Department of Physics,
Department of Astronomy,
Box 351580,
University of Washington,
Seattle, Washington, 98195, USA
e-mail: stubbs@radon.astro.washington.edu

Cloning, dignity and ethical revisionism

Sir — Several texts published in *Nature*, notably a leading article devoted to the opinion of the French Consultative Ethics Committee (*Nature* 387, 321 & 324; 1997) and John Harris's letter (*Nature* 387, 754; 1997), discuss the argument that application to man of asexual reproduction and cloning represents an affront to human dignity.

However, the two main arguments I set out in my Commentary (*Nature* 386, 119; 1997), and which were expanded in the French ethics committee opinion, were not accurately reported.

One of the components of human dignity is undoubtedly autonomy, the indeterminability of the individual with respect to external human will. No man or woman on Earth exists exactly as another has imagined, wished or created.

The birth of an infant by asexual reproduction would lead to a new category of people whose bodily form and genetic make-up would be exactly as decided by other humans. This would lead to the establishment of an entirely new type of relationship between the 'created' and the 'creator', which has obvious implications for human dignity.

Harris contests the validity of arguments based on the Kantian principle. But Kant did not say that respect for human dignity requires that an individual is *never* used as a

means, but that an individual must never be used *exclusively* as a means. The word 'exclusively' makes all the difference between idle talk and one of the fundamental principles of modern bioethical thought.

The workman is indeed the means for doing work, the person who donates an organ is the means for saving the patient, but they are never exclusively a 'means', but also ends in themselves.

The creation of human embryos exclusively as a means, uniquely as a source of therapeutic material, would therefore seem in contradiction of Kant's principle, whose universality is far superior to that which Harris dismisses by omitting the word 'exclusively'.

In reality, the debate is about the status of the human embryo and its rights as a human individual, and the answers to this question differ greatly both between and within nations. In general, however, all those who would legitimize *de novo* creation of human embryos for research or preparation of therapeutic material base their position on their belief that the embryo is not a human individual, without calling Kant's principle into question.

Is Harris announcing the emergence of a revisionist tendency in bioethical thinking?

Axel Kahn

Institut Cochin de Génétique Moléculaire,
24 Rue du Faubourg Saint-Jacques,
F-75014 Paris,
France
e-mail: kahn@icgm.cochin.inserm.fr

Salmonella or Smithella?

Sir — The history of science contains many examples of the subtitle of your recent leading article, "Authorship of a scientific paper is a privilege that is all too easily abused"¹. The consequences can be far-reaching for later honour and glory.

The salmonellae were named in 1900 undeservedly after the US veterinarian Daniel Elmer Salmon (1850–1914), director of the Bureau of Animal Industry of the US Department of Agriculture.

These pathogenic germs should in fairness be called smithellae, because their first member, called today *Salmonella choleraesuis*, was discovered by the greatest American pioneer of microbiology, Theobald Smith^{2,3} (1859–1934), who later (in 1893) also found the microbe that causes Texas cattle fever. But the first author of the paper "The bacterium of swine-plague"⁴ was Salmon, who had not taken part in the research at all and was only the superior of Smith, who was named as the second author. Unfortunately the designation *Salmonella* is so firmly rooted that it would be impossible to change it.

Friedrich Katscher

Mariahilfer Str. 133,
A-1150 Vienna, Austria

1. *Nature* 387, 831 (1997).
2. Dolman, C. E., *Theobald Smith*, in *Dictionary of Scientific Biography* (Ed. Gillespie, C. C.) 12, 480–486 (Charles Scribner's Sons, New York, 1980).
3. Bibel, D. J. *Milestones in Immunology*, 31–32 (Springer, Berlin, 1988).
4. *Amer. Monthly Microsc. J.* 7, 204–205 (1886).

Fat rats and carcinogenesis screening

It is poor science to use inappropriate strains of rodents in badly designed and inefficient experiments for testing the safety of chemicals. It is also costly and unethical because greater numbers of laboratory animals are needed.

Michael F. W. Festing

Toxicologists do not seem to understand the importance of using isogenic strains of laboratory animals in the toxicity testing of chemicals, nor do they appreciate the importance to their testing procedures of genetic variation in both humans and animals. In most other biological disciplines, strain differences are used as a powerful research tool. But it is an extraordinary fact that virtually all toxicity testing is done using a single outbred stock of genetically undefined, heterogeneous animals, totally ignoring the many advantages of using isogenic strains.

Toxicologists are now in trouble. Their laboratory rats and mice are "fat, frail and dying young"¹, sometimes before the tests for long-term toxicity are completed. The problems arise from 'improved' diets and environments, but are compounded by the use of genetically labile outbred stocks that have become heavier as a result of long-term selection for faster growth, and also by a failure to develop more suitable long-lived 'isogenic' strains (inbred strains and the F₁ hybrid between two such strains).

These problems are serious. Carcinogenesis screening experiments probably cost more than US\$250 million a year and involve more than 70,000 animals². The development of several transgenic mouse strains that are highly susceptible to carcinogens³, international harmonization and attempts to reduce the costs and use of animals for carcinogenesis screening have led to the suggestion that toxicity tests could be done using only the rat, with one or more highly sensitive transgenic strains of mice to provide additional information. Rat studies must now be redesigned using a battery of isogenic rat strains if *in vivo* tests are to be continued for the foreseeable future. I suspect

that the poor quality of

present experimental designs is leading to inaccuracies that make it extremely difficult to validate alternative *in vitro* and short-term tests.

Fat-rat syndrome

Every year many chemicals are tested to assess the conditions under which they can be safely used. Following extensive *in vitro* and *in vivo* tests, a final step often involves an 18–24-month carcinogenesis study in mice and rats. At least 50 per cent of the animals should survive to the end of the test period for the study to be acceptable. Improvements in diet, husbandry and the animal-house environment, as well as the elimination of infection⁴, have resulted in fatter rats with more cancer and heart disease and a shorter lifespan. Although isogenic strains have also been affected, the most serious problems are found among some of the outbred stocks of Sprague–Dawley and Wistar rats, selected for fast growth rate.

Historically, outbred stocks have been chosen according to availability. Once a company has used a particular stock, it is reluctant to change because it assumes that historical data are useful in the interpretation of test results. Problems associated with poor initial choice are further exacerbated by the apparent lack of understanding of the properties of the outbred stocks that are most widely used.

Outbred stocks have serious disadvantages when used in controlled experiments. Stocks change over time as a result of mutation, genetic drift and/or directional selection, and sometimes by genetic contamination from another stock. Daughter colonies will differ genetically from their parents as a result of genetic sampling and selection, so eventually each colony becomes genetically distinct, even though they may all have the same name (for example 'Wistar'). Little is known about the genetics of any of these stocks. There are no genetic tests that can be used to distinguish between Wistar and Sprague–Dawley, the two most widely used rat stocks. The absence of genetic quality control and environmental changes make the value of historical control data for outbred stocks highly questionable⁵.

In contrast, isogenic strains are produced by many generations of brother/sister mating and can be regarded as immortal clones of genetically identical individuals. Treated and control groups are exactly matched genetically, usually leading to higher statistical power. Any genetic drift due to new mutations is orders of magnitude slower than that

occurring in outbred stocks. Colonies with the same name are genetically virtually identical, so historical data can be gleaned from the literature, subject to environmental influences. Finally, there is an important precedent. More than 450 chemicals have been screened in the US National Toxicology Program carcinogenesis bioassay using isogenic F344 rats and B6C3 F₁ hybrid mice⁶.

Why outbred stocks are used

Inbred strains of mice and rats have been available for well over half a century, and are now widely used in most scientific disciplines. Several Nobel prizes have been awarded for work that would have been impossible without them. Yet, in toxicology, there seems to be an almost irrational prejudice⁷ against their use in both screening⁸ and research⁹. Screening has nearly always been done using outbred animals, so it is the path of least resistance for regulators and the pharmaceutical industry to continue in the same way. Although there have been few written attempts to justify the use of such stocks, two arguments (mostly expressed verbally) are used to do so.

The first usually claims that "we wish to model humans, humans are outbred, therefore we should use outbred laboratory animals". This argument has the same logical structure as Woody Allen's claim that "God is good, I am good, therefore I am God", and makes about as much sense as a claim that "we wish to model humans (true), humans have no tail (true), therefore we should use animals with no tail". There is no need for a model to be exactly like its target in every respect. If it were, it would not be a model.

A variant of the same argument suggests that it should be easier to 'extrapolate' the results of an experiment to genetically varied humans if it is based on genetically varied outbred animals. However, even if this were true (which it is not), the first requirement must be to design a good experiment with a low chance of false-negative or false-positive results. This cannot be achieved by the use of a single outbred stock. There is no point in trying to make it easier to extrapolate results if they are wrong in the first place.

In any case, 'extrapolation' does not adequately describe the way in which toxicological data are interpreted. Are the results of the Ames test, conducted on the genetically modified bacterium *Salmonella*, 'extrapolated' to humans? Will the results of studies with transgenic mice be 'extrapolated' to non-transgenic humans? If so, there should be no difficulty in extrapolating from inbred



rats. If not, then how are such data used? I think that Popper's deductive approach¹⁰ corresponds more closely with reality, as follows.

Let us start with the premise that it is impossible to prove that any compound is safe, but it is sometimes possible to show that it is unsafe. If rigorous testing fails to show that it is unsafe (possibly under specified conditions of use), then it is provisionally accepted as being safe. Carcinogenesis screening involves a series of testable hypotheses chosen to be relevant to human safety.

The first might be that the compound is not genotoxic because, if it is, it is more likely to be a human carcinogen. This is explored using a battery of short-term tests such as the Ames test. Failure to detect genotoxicity would lead to the compound provisionally being accepted as non-genotoxic, although a more sensitive test may later show genotoxicity. The next hypothesis might be that the compound is not a mammalian carcinogen. This can be tested experimentally using rodents, possibly including the transgenic strains. If the compound causes cancer in rodents, then there is a high probability that it may also cause cancer in humans. There is no 'extrapolation' between species. It is decided before the experiment is started that a positive result implies that the compound is a mammalian carcinogen and a negative result that it will provisionally be accepted as a non-carcinogen.

The level of confidence in the results depends on the rigour of the screening process. Experiments need to be both relevant and well designed if the screening is to be acceptable. The only extrapolation is the within-species mathematical extrapolation of the dose-response curve to estimate safe human exposure, and this is only as good as the biological data (rubbish in, rubbish out). Finally, the hypothesis that the compound is not a human carcinogen even if it is a rodent carcinogen might be tested, although this is difficult and is rarely attempted.

The second argument employed to justify the use of outbred stocks is that "... it is more correct to test on a random-bred stock on the grounds that it is more likely that at least a few individuals will respond to the administration of an active agent in a group which is genetically heterogeneous"¹¹. Unfortunately, this shows a lack of understanding of the properties of outbred stocks, and of experimental design. It is highly desirable to have a test population with a wide range of susceptible phenotypes. However, the use of a single outbred stock is a very inefficient way of achieving that objective because phenotypic variation in susceptibility within a population is usually much less than the variation between different populations. So it makes more sense to sample from more than one population.

Toxicologists are now in trouble... the poor quality of experimental designs is leading to inaccuracies that make it extremely difficult to validate alternative tests

Moreover, the main requirement for a 'good' controlled experiment is that the treated and control groups should, as far as possible, be identical at the start of the experiment. Uncontrolled heterogeneity for genes associated with spontaneous disease, body weight and physiological processes makes it impossible to have genetically matched treated and control samples when using outbred stocks. The resulting 'noise' can obscure biologically important treatment effects, or even produce false-positive (or false-negative) results. Toxicologists try to match the treatment groups for age, body weight and health status but mistakenly seem to think that genetic heterogeneity is different, and is a 'good thing'.

Different and better strains

In some cases a strain can be totally resistant to a compound that causes a high incidence of tumours in other strains. For example, in one study diethylstilbestrol (DES) reduced the number of tumours in outbred Sprague-Dawley rats, but caused tumours in more than 70 per cent of inbred ACI rats¹². DES would have been classified as a non-carcinogen had only Sprague-Dawley rats been used.

By using more than one strain, the range of susceptibility phenotypes can usually be increased without increasing the total number of animals. Assuming 48–50 rats in each group, they could all be the same strain, or there could be 25 rats of 2 different strains, 12 rats of 4 strains, 10 rats of 5 strains or even 1 rat of 50 strains (at least in theory, although this would not be practical). A design using, say, 16 rats of 3 strains would provide reasonable assurance that a totally resistant strain had not been chosen. Treated and control groups would be exactly matched genetically at the start of the experiment, and total sample size in each group would be unchanged. Toxicologists are often sceptical that such a design is statistically valid, but it is. On average, for a given number of animals, a multi-strain design has higher statistical power than either the use of a single outbred stock¹³ or a single inbred strain¹⁴, with power increasing with the number of strains. A study of the reasons for any strain differences that emerge could also provide information relevant to estimating human risk⁸.

Ideally, the isogenic strains used will have a long lifespan and an acceptable spectrum of

spontaneous pathology. Yet, despite the many billions of dollars spent on carcinogenesis screening, and all the problems with the fat-rat syndrome, no attempt seems to have been made by the pharmaceutical industry to find such strains.

Experimental gerontologists have fewer resources than toxicologists, yet the US National Institute of Aging has been evaluating several isogenic strains of mice and rats for gerontological studies¹⁵. Rat strains F344, BN/Rij and the F344BN F₁ hybrid have quoted mean lifespans of 103, 130 and 146 weeks, respectively, and have reasonably acceptable spontaneous pathology¹⁶. More long-lived strains particularly suitable for carcinogenesis screening could be found and used by the whole pharmaceutical industry as part of the international harmonization currently being discussed. Genetic change can be eliminated by maintaining banks of frozen embryos¹⁷, and standardized methods of genetic quality control, which are already well developed, could be applied routinely to ensure genetic integrity¹⁸.

Eventually, it should be possible to develop *in vitro* tests that are sufficiently accurate to make animal testing unnecessary. Validation would be easier if the quality of the animal data could be improved. In the meantime, the suggestion that testing should be based on a conventional carcinogenesis screen in rats plus other data, which might include short-term tests in transgenic mice, is gathering support.

With the welcome success of international harmonization of testing methods, toxicologists and governmental regulators must now improve the design of rat carcinogenesis screening experiments using a suitable test panel of isogenic rat strains. □

Michael F. W. Festing is at the Medical Research Council Toxicology Unit, University of Leicester, PO Box 138, Leicester LE1 9HN, UK.

e-mail: mfwf1@leicester.ac.uk

1. Nohynek, G. J. *et al.* *Hum. Exp. Toxicol.* **12**, 87–98 (1993).
2. McAuslane, J. A. N., Lumley, C. E. & Walker, R. S. *Hum. Exp. Toxicol.* **10**, 205–213 (1991).
3. Tennant, R. W., Spalding, J. & French, J. E. *Mutat. Res.* **365**, 119–127 (1996).
4. Rao, G. N., Haseman, J. K. & Edmondson, J. *Lab. Anim. Sci.* **39**, 389–393 (1989).
5. Roe, F. J. C. *Lab. Anim.* **28**, 148–154 (1994).
6. Haseman, J. K. & Seilkop, S. K. *Fundam. Appl. Toxicol.* **19**, 207–213 (1992).
7. Abelson, P. H. *Science* **255**, 141 (1992).
8. Festing, M. F. W. *Environ. Health Perspect.* **103**, 44–52 (1995).
9. Festing, M. F. W. *Toxicol. Appl. Pharmacol.* **102**, 197–204 (1990).
10. Popper, K. R. *Objective Knowledge* (Clarendon, Oxford, 1972).
11. Arcos, J. C., Argus, M. F. & Wolf, G. (eds) *Chemical Induction of Cancer* (Academic, London, 1968).
12. Shellabarger, C. J., Stone, J. P. & Holtzman, S. J. *Natl Cancer Inst.* **61**, 1505–1508 (1978).
13. Haseman, J. K. & Hoel, D. G. *J. Toxicol. Environ. Health* **5**, 89–101 (1979).
14. Felton, R. P. & Gaylor, D. W. *J. Toxicol. Environ. Health* **26**, 399–411 (1989).
15. Sprott, R. L. *Aging Clin. Exp. Res.* **5**, 249–252 (1993).
16. Bronson, R. T. & Lipman, R. D. *Aging Clin. Exp. Res.* **5**, 253–257 (1993).
17. Nomura, T. & Potkay, S. *Institute of Laboratory Animal Resources News* **33**, 42–44 (1991).
18. Festing, M. F. W. *Aging Clin. Exp. Res.* **5**, 309–315 (1993).

Seeking wisdom in innate immunity

Douglas T. Fearon

Obsolescent — such was once the view of the innate immune system. That view is itself out of date. But a fresh angle of attack on the ways in which innate immunity is essential to the function of acquired immunity comes from study of the ancient defence systems of invertebrates.

The crowning accomplishment of vertebrate immunity is the acquired immune system. It produces trillions of clones of B and T lymphocytes, each expressing a unique antigen receptor, ensuring that all microorganisms will have some structure, or antigen, that can be immunologically recognized (see box). To achieve this diversity, developing lymphocytes somatically create unique antigen receptors by recombining segments of genes and randomly adding or subtracting nucleotides at sites of recombination¹. An antigen selects only those lymphocytes that have receptors of sufficient affinity for growth and differentiation.

But this genetically economical solution to producing enormous protein diversity comes with a price. It uncouples acquired immune recognition from the wisdom gained during evolution of how to cope with hostile microbes because, for the most part, the antigen receptors are not encoded in the germ line. With no information from the past, these receptors can evaluate only

antigenic structure and not its biological correlates.

Where can we find the biological wisdom that could instruct acquired immunity? In their paper on page 394 of this issue², Medzhitov, Preston-Hurlburt and Janeway propose that the search for such wisdom should be made in innate immunity, and that the clues that direct these searches can be found in ancient defence systems of invertebrates. They back up this suggestion by finding a human homologue of the *Drosophila* Toll protein, a receptor involved in innate immunity of this invertebrate, that has the potential for potentiating the T-lymphocyte response to antigen.

Acquired immunity only appeared with the evolution of vertebrates, but before then innate immune systems had been successfully defending invertebrates and plants against microbial infections for hundreds of millions of years. The germline-encoded receptors of innate systems are relatively limited in diversity and unable to make fine distinc-

tions between closely related structures. Nevertheless, they can recognize certain chemical features shared by groups of microorganisms but not by the host, such as lipopolysaccharide of Gram-negative bacterial cell walls. This capability enables innate immunity to detect the presence of an infection, if not the precise cause — a biological rather than a structural distinction.

Despite its evolutionary success, innate immunity has been treated with condescension by immunologists. It has been considered a stopgap measure, a temporary expedient for host defence, buying time until acquired immunity took over. It lacked the elegance of the genetic-recombination mechanism to generate specificity. It had no 'memory', because first and second exposures to a microbial substance elicited similar responses (curiously, this attitude neglected the evolutionary memory that allowed the rapid first response). Even worse, its overreaction to bacterial products in some circumstances caused lethal shock, suggesting that its inflammatory and cytotoxic capabilities were out of proportion to its judgement and discretion. In short, innate immunity was unsophisticated, unintelligent, indiscreet and obsolescent.

In the 1989 Cold Spring Harbor Symposium on immune recognition, Janeway acknowledged this hubris of 'acquired immunologists' when he referred to their "dirty little secret", the need to contaminate pure antigens with adjuvants to induce acquired immune responses³. Adjuvants are unpleasant concoctions of oils, salts, killed

Innate and acquired immunity

Immunity to infection is mediated by two general systems, innate (or natural) and acquired (or adaptive). Innate systems use germline-encoded proteins to identify microbial substances. For example, mammalian macrophages have mannose, lipopolysaccharide and scavenger receptors for identifying components of yeast and bacterial cell walls. Plasma proteins, such as the complement cascade, C-reactive protein, and the mannose-binding protein of the collectin family, also recognize microbial carbohydrates. These cellular and soluble proteins pre-exist, or are rapidly induced within hours of infection. They provide immediate defence against microorganisms by

directing the engulfment and breakdown of microbial cells, and the secretion of cytokines and chemokines which enhance antibacterial and antiviral cellular responses. Given that the receptors are 'hard-wired' in their ability to detect infectious organisms, innate immunity has the disadvantage of genetic inflexibility, meaning that microbes can evade surveillance through mutation.

Acquired immunity is the system of B and T lymphocytes, each of which expresses a distinct antigen receptor that has arisen, not through a process of evolution, but by somatic gene rearrangement. The generation of specificities is

without regard to targets, and leads to the creation not only of potentially microbe-specific receptors but also to self-reactivity — the self/non-self problem of acquired immunity. These undesired clones must be eliminated during their development, or perhaps hindered in their activation by self-antigen, to prevent autoimmunity.

The residual lymphocyte clones represent the pool from which those with microbial reactivity can be selected. However, on first exposure to an infectious organism, the frequency of such clones is initially too low, perhaps as few as 1 in 10,000–100,000, for a protective response. During the ensuing 1–2 weeks these

clones replicate, increasing in number by as much as 1,000 times, and mature into cells that defend the host by secreting antibodies, activating macrophages and killing virally infected cells. When the infection is cleared, many of the antigen-specific B and T lymphocytes remain as 'memory' cells which can mount a rapid response if the same microorganism is encountered again. It is this characteristic that accounts for the designation 'acquired immunity'. In effect, this system compresses evolution into two weeks, allowing the selection of specific antigen receptors to keep pace with the selection of variant microorganisms, overcoming the deficiencies of innate immunity. **D.T.F.**

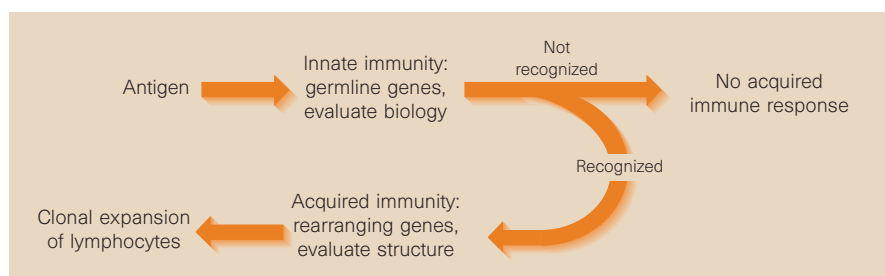


Figure 1 Integration of innate and acquired immune recognition. In a primary immune response, antigen is first subjected to biological evaluation by innate immunity for its possible microbial origin. If considered as such, antigen is presented in a more immunogenic form to the acquired immune system, causing the clonal expansion of lymphocytes that recognize antigen with a high degree of structural specificity. B lymphocytes then produce antibodies to eliminate the antigen-carrying invader, such as a bacterium. Or different types of T lymphocyte activate macrophages to destroy intracellular microbes, or themselves kill cells infected, for instance, by a virus.

bacteria, bacterial products and so on, none of which has anything to do with the interaction of antigen and antigen receptors on lymphocytes, but all of which magnify the acquired immune response. He suggested that adjuvants had this effect because innate immunity recognized them as microbial, and the admixed antigen became guilty by association. This judgement was transmitted to the acquired immune system to induce a vigorous response to the antigen.

We now know of several means by which innate immunity communicates its biological evaluation of an antigen to acquired immunity: the secretion of cytokines by macrophages and natural killer cells⁴; the attachment of a complement protein to antigen⁵; the preferential uptake of microbial antigen by lectin receptors on cells that specialize in presenting antigen to lymphocytes⁶; the maintenance of antigen-reactive T lymphocytes⁷; and the expression of the membrane proteins B7.1 and B7.2 on the antigen-presenting cells⁸. The B7.1 and B7.2 proteins, which can be induced by lipopolysaccharide, interact with receptors on T lymphocytes to deliver a 'second signal', the first being that of the antigen receptor. Because two signals are required for lymphocyte activation, Medzhitov and colleagues' finding² that the human homologue of Toll induces B7.1 on macrophages places it in the communication network between innate and acquired immunity.

Interpreting the 'second signal' as a link connecting innate and acquired immunity heralds a fundamental shift in emphasis in immunology. This signal was originally proposed as a solution to the problem of self/non-self discrimination. With the second signal now being related to the biased recognition of microbial antigens, immunology moves away from self/non-self as a unifying concept towards the new dichotomous classifications, infectious/non-infectious^{3,9} and dangerous/not dangerous¹⁰. These distinctions necessarily depend on the capabilities of innate immunity, and seem to be closer to the essence of the immune

response, which is protection from infection.

If we accept the principle that the biological wisdom of innate immunity must be integrated with the structural specificity of acquired immunity (Fig. 1), how do we identify proteins that mediate this process? Medzhitov *et al.*² provide an approach that should be generally applicable: search the ever more complete library of the human genome for mammalian homologues of immune proteins in invertebrates. Their choice of the search term, *Drosophila* Toll, was especially apt because it activates Dorsal,

a gene transcription factor with homologues in the mammalian NF- κ B system which is involved in so many innate and acquired immune reactions¹¹.

Finally, what will we do with these newly identified proteins? We will understand better what makes an antigen immunogenic, both self and foreign. This will lead to more effective therapies for autoimmunity and vaccines for infectious diseases and tumours. By then, even the most sceptical among us will be convinced of the wisdom of innate immunity. □

Douglas T. Fearon is in the University of Cambridge School of Clinical Medicine, Department of Medicine, Wellcome Trust Immunology Unit, Hills Road, Cambridge CB2 2SP, UK.

1. Ramsden, D. A., van Gent, D. C. & Gellert, M. *Curr. Opin. Immunol.* **9**, 114–120 (1997).
2. Medzhitov, R., Preston-Hurlburt, P. & Janeway, C. A. Jr *Nature* **388**, 394–397 (1997).
3. Janeway, C. A. Jr *Cold Spring Harb. Symp. Quant. Biol.* **54**, 1–13 (1989).
4. Unanue, E. R. *Curr. Opin. Immunol.* **9**, 35–43 (1997).
5. Dempsey, P. W., Allison, M. E., Akkaraju, S., Goodnow, C. C. & Fearon, D. T. *Science* **271**, 348–350 (1996).
6. Jiang, W. *et al. Nature* **375**, 151–155 (1995).
7. Vella, A. T., McCormack, J. E., Linsley, P. S., Kappler, J. W. & Marrack, P. *Immunity* **2**, 261–279 (1995).
8. Lenschow, D. J., Walunas, T. L. & Bluestone, J. A. *Annu. Rev. Immunol.* **14**, 233–258 (1996).
9. Fearon, D. T. & Locksley, R. M. *Science* **272**, 50–53 (1996).
10. Matzinger, P. *Annu. Rev. Immunol.* **12**, 991–1045 (1994).
11. Baeuerle, P. A. & Baltimore, D. *Cell* **57**, 13–20 (1996).

Solar physics

Gravity waves with a new spin

Douglas Gough

The Sun loses angular momentum to the solar wind via its magnetic field, slowing down the rotation of the outer, convective layers. The core was once thought to be weakly coupled to the outer layers, rotating up to ten or more times faster than the convection zone, but helioseismology has shown that much of the radiative interior rotates at a similar rate to the surface. What couples the two regions so strongly? In two recent papers^{1,2} wave transport is invoked: gravity waves (similar to internal ocean waves, whose restoring force is buoyancy), generated near the interface between the convective and radiative regions, transport retrograde angular momentum into the interior, thereby spinning it down. Previous estimates³ had cast doubt on the efficacy with which the transport can take place. So what has changed?

The Sun can be divided into two zones, the upper layer where heat transport is by convection, and a lower, stable layer where it is radiative. The angular velocity of the surface decreases with latitude⁴, the value at the poles being about 70 per cent of that at the equator. Helioseismology — the study of small oscillations in the shape of the Sun — can probe conditions below the surface, and has shown that variation persists down to the base of the

convection zone⁵. Most of the radiative interior rotates almost uniformly at an intermediate rate. The radiative zone is separated from the convection zone by a thin shear layer, called the tachocline⁶, whose thickness is about 5 per cent of the Sun's radius⁷.

Turbulence can redistribute angular momentum within the convective zone in less than a year, so we can assume that this zone adjusts itself to its own will, and then ask why the interior rotates as it does. There are two main questions. Why does the interior rotate almost rigidly? And why does it rotate at the rate it does?

A natural agent for stiffening the interior is a magnetic field. Just 10 μ G (10^{-9} T), a value comparable to the interstellar field, can redistribute enough angular momentum over the lifetime of the Sun for the core to rotate rigidly. But popular opinion is that such a field should penetrate into the convection zone, and thereby transmit the differential rotation of the outer envelope into the core. (Indeed, a modicum of penetration at high latitudes might cause the apparently slow rotation of the innermost core⁸.) Alternatively, as a consequence of the circulation in the tachocline (see Fig. 1), the global field could be almost wholly contained in the

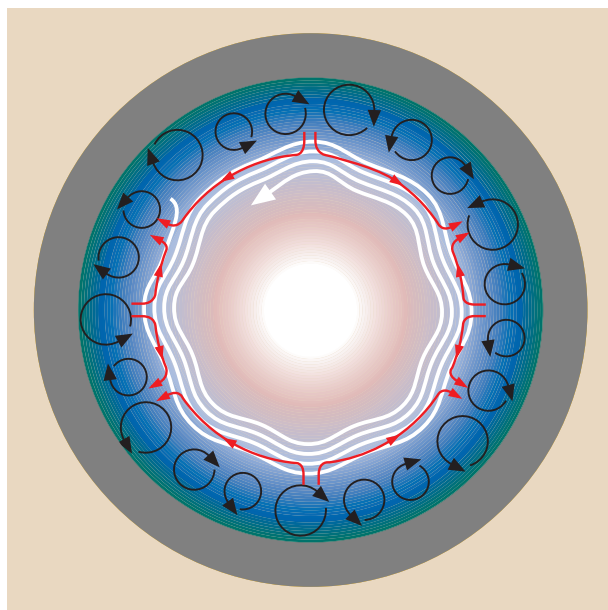


Figure 1 A section through the Sun. The largest eddies, at the base of the convection zone, generate waves which spiral through the thin 'tachocline' into the interior (the radiative zone). Some of these waves, on spirals roughly 100 times tighter than this one, carry angular momentum inwards, and may be responsible for coupling the rotation of outer and inner regions.

radiative interior, leaking out, perhaps, only where the angular velocities of the radiative and convection zones coincide, where there is negligible rotational shear across the tachocline. That occurs at latitudes of about 30° , approximately the maximum extent of the sunspot belt, and may be related to the onset of each new sunspot cycle.

Another mechanism invoked for inducing uniform rotation is turbulence. Spiegel and Zahn⁶ assumed two-dimensional turbulence in the tachocline, isotropic on horizontal surfaces. The associated stresses would lead to latitudinally invariant rotation at depth, but not to the near uniformity with radius that is observed. However, as in the case of the Earth's stratosphere, the turbulence is likely to be anisotropic, leading to latitudinal shear. It therefore does not provide an explanation for the helioseismic data.

Let us accept that the radiative interior does rotate rigidly, and ask at what rate that should be. Because the convection zone can adjust almost immediately to the rotation of the interior, the net torque across the tachocline must be very small. Given that, two estimates of the angular velocity Ω_0 of the radiative zone have been made. One⁹ assumes a linear relation between stress (force) and strain (distortion) across the tachocline, as would occur in three-dimensional turbulence, which leads to $\Omega_0 = 0.96\Omega_c$, where Ω_c is the equatorial angular velocity in the convection zone. The other is the two-dimensional Spiegel-Zahn model⁶, with no turbulent transmission of stress across the tachocline, which yields $\Omega_0 = 0.90\Omega_c$. Helioseismic observation⁵ yields an intermediate value, $\Omega_0 = 0.93\Omega_c$. So some horizontal stress is being transmitted radially. The hypothesis discussed in refs 1 and 2 is that it is accomplished not by shear turbulence, but by gravity waves.

Waves generated by the convection must satisfy two conditions in order to control the

rotation of the radiative interior. They must be generated with enough amplitude to transport the requisite angular momentum; and they must deposit their angular momentum before returning to the convection zone, yet not before penetrating far into the radiative interior. The first condition is satisfied by gravity waves with the same frequency and horizontal wavelength as the convection^{2,10} (which are roughly 2×10^{-7} Hz and $0.1R_\odot$, respectively, where $R_\odot$ is the solar radius). However, these resonant waves penetrate less than $10^{-5}R_\odot$ beneath the convection zone³.

The depth to which the waves penetrate varies as the fourth power of frequency, and as the cube of the horizontal wavelength. So Zahn, Talon and Matias² can more than overcome this factor of 10^5 by assuming that the resonant frequency is 2π times that adopted previously^{3,10}, and by concentrating on waves with ten times the resonant wavelength. Such waves can reach the core. But can waves so far from resonance be generated with enough amplitude? There is so much uncertainty in the properties of solar convection that the possibility cannot be ruled out³.

What are the consequences of the wave-transport hypothesis? Prograde and retrograde modes are generated equally when viewed in a frame rotating with the source, yet angular momentum can be transported only if an appropriate asymmetry is generated. This could come about if the angular velocity in the Sun's radiative zone changes with depth, as might be expected during spin-down. Then prograde and retrograde modes have different frequencies in a frame rotating locally with the Sun, and so they dissipate differently.

Zahn, Talon and Matias² concentrate on the early stages of spin-down, when the rotation might have changed rapidly enough with depth for 'critical-layer absorption'¹¹: a wave is absorbed completely as it approaches

the level where its angular phase velocity matches the angular velocity of the Sun. This can happen only to prograde modes (assuming the Sun to be rotating more rapidly at greater depths), so there is very effective angular-momentum transfer. Kumar and Quataert¹ consider only noncritical dissipation, which is weaker, but still strong enough to transfer a lot of angular momentum. Again, prograde modes are dissipated preferentially.

In both papers, it is assumed that the transport is in such a direction as to reduce the shear, leading eventually to uniform rotation of the radiative interior, as is observed⁵. But the preferred dissipation of prograde modes would tend to speed up the interior rather than slow it down³. Only at greater depths can the effect of the wave-induced stress be reversed, where the retrograde-prograde amplitude ratio becomes high, because the prograde modes have already been preferentially absorbed. (This is the process that drives the quasibiennial oscillation of the Earth's atmosphere¹², and it has been demonstrated in the laboratory by Plumb and McEwan¹³.) So perhaps the more deeply penetrating retrograde modes slow the solar core.

The new work^{1,2} has highlighted the importance of off-resonant waves, which may play an essential role in determining variations of rotation in the Sun. They may also carry a lot of heat downwards (thereby explaining in part the sound-speed discrepancy between the Sun and standard models¹⁴), homogenize the helium that has settled beneath the convection zone, and deplete lithium and beryllium in the solar photosphere¹⁵. □

Douglas Gough is at the Institute of Astronomy, The Observatories, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK, and is a member of the Department of Applied Mathematics and Theoretical Physics, University of Cambridge.

- Kumar, P. & Quataert, E. *J. Astrophys. J.* **475**, L143–L146 (1997).
- Zahn, J.-P., Talon, S. & Matias, J. *Astron. Astrophys.* **322**, 320–328 (1997).
- Gough, D. O. in *The Energy Balance and Hydrodynamics of the Solar Chromosphere and Corona* (eds Bonnet, R.-M. & Delache, P.) 3–36 (de Bussac, Clermont-Ferrand, 1977).
- Snodgrass, H. B. *Solar Phys.* **94**, 13–31 (1984).
- Thompson, M. J. *et al. Science* **272**, 1300–1305 (1996).
- Spiegel, E. A. & Zahn, J.-P. *Astron. Astrophys.* **265**, 106–114 (1992).
- Kosovichev, A. G. *Astrophys. J.* **469**, L61–L64 (1996).
- Elsworth, Y. *et al. Nature* **376**, 699–672 (1995).
- Gough, D. O. in *Future Missions in Solar, Heliospheric and Space Plasma Physics* (eds Rolfe, E. J. & Battrick, B.) 183–197 (ESA SP-235, Noordwijk, 1985).
- Press, W. H. *Astrophys. J.* **245**, 286–303 (1981).
- Booker, J. R. & Bretherton, F. P. *J. Fluid Mech.* **27**, 513–539 (1967).
- McIntyre, M. E. in *The Solar Engine and its Influence on the Terrestrial Atmosphere and Climate* (ed. Nesme-Ribes, E.) 293–320 (Springer, Berlin, 1994).
- Plumb, R. A. & McEwan, A. D. *J. Atmos. Sci.* **35**, 1827–1839 (1978).
- Gough, D. O. *et al. Science* **272**, 1296–1300 (1996).
- Gough, D. O. in *Solar-Terrestrial Relationships and the Earth Environment in the Last Millennium* (ed. Castagnoli-Cini, G.) 90–132 (Soc. Italiana di Fisica, Bologna, 1988).

Palaeoanthropology

Fossil muzzles and other puzzles

Meike Köhler and Salvador Moyà-Solà

Some time between the late Oligocene and early Miocene, 28–22 million years (Myr) ago, the initial evolutionary radiation which led to the living catarrhines took place. The catarrhines are a group composed of the Old World monkeys and the hominoids, which today are respectively represented by the cercopithecine and colobine subfamilies of monkey, and the lesser apes, great apes and humans (Fig. 1). On page 368 of this issue¹, Benefit and McCrossin provide an account of the “Earliest known Old World monkey skull”, and thereby add further fuel for debate about catarrhine evolutionary history.

Several attempts have been made to construct a reliable evolutionary framework that can both accommodate the increasing number of fossil catarrhine taxa and yet still permit the recognition of the earliest ‘stem’ members of extant Old World primates. The outcome, however, greatly depends upon the assessment of primitive (ancestral) and derived (modern) traits. No consensus has yet been achieved as to what, exactly, are the ancestral states for the common ancestor of monkeys and apes. According to the most commonly accepted hypothesis, the cranial morphology of large-eyed, round-headed and short-

snouted primates, such as the hylobatids (the lesser apes — the gibbon and siamang) and colobines², closely approximates to the primitive catarrhine pattern³. But some authors⁴ believe that the rather long-snouted Oligocene catarrhine *Aegyptopithecus*, best known from the Fayum of Egypt and dated to about 34 Myr ago, provides a better model for the primitive condition.

It is in this context that Benefit and McCrossin’s description of a beautifully preserved cranium of the middle Miocene Old World monkey *Victoriapithecus*, from Maboko Island, Kenya, promises to be of critical importance. The specimen is estimated to be roughly 15 Myr old; it is the only complete and undistorted catarrhine cranium known from between the middle Oligocene and late Miocene, and thus offers the first evidence of cranial morphology from this 25-Myr interval of time. This specimen may end the debate about the character states that are hypothesized to have existed in the common ancestor of monkeys and apes.

Benefit and McCrossin argue that the new specimen confirms that the primitive character states for catarrhines include a low cranial vault, linear facial profile, moderately long snout, and a slight degree of vertical deflec-

tion in the face (‘klinorhynch’), thus matching the cranial morphology seen in the older *Aegyptopithecus*. The main features of *Victoriapithecus* challenge the notion that the ancestral condition for modern Old World primates is most closely matched by the round-headed and short-snouted hylobatids and colobines³.

Many evolutionary questions can now be tackled. Benefit and McCrossin choose to address the issue of hominoid phylogeny and they provide a list of discrete cranial characters found in *Victoriapithecus* that they also consider to be primitive for hominoids. They argue that some of these character states are also present in the Miocene apes, *Dryopithecus* from Europe and *Sivapithecus* from Asia, as well as the extant orang-utan from Southeast Asia — thus calling into question proposals that there is a close affinity between these fossil apes and the great apes and humans.

However, shared features are not strictly limited to the retention of primitive states or derived states traced to descent from a common ancestor; shared features can also evolve independently because of similar functional demands or structural constraints. Take, for example, the case of the frontal-trigone morphology of *Victoriapithecus*. Benefit and McCrossin consider this structure to be primitive, but it simply reflects the size relation between the chewing musculature and the braincase. So it may result either from small brain size or from large muscles as an allometric effect (during growth, the mass of the chewing muscles increases more than does the size of the braincase, so that in large individuals additional crests are needed to provide sufficient surface area for muscle attachment⁵). The same applies to tall and narrow orbits, which the authors also regard as primitive. These might result from such unexpected factors as the constraints imposed by the diameter of the birth canal, as suggested for the unique interorbital condition of the New World squirrel monkey⁶, and thus could have been derived independently several times.

Comparing overall skull structure, however, we think that this new finding underscores the clear differences between the cranial patterns of primitive early Miocene African stem hominoids (such as *Afropithecus*) and those of the late Miocene Eurasian hominoids (such as *Sivapithecus* and *Dryopithecus*). The new discovery shows that *Aegyptopithecus* is very like the ancestral catarrhine in cranial morphology, so it also confirms assertions that the facial morphology of *Afropithecus* is basically primitive because it resembles that of *Aegyptopithecus*⁷. However, unlike *Afropithecus*, the late Miocene Eurasian hominoids have considerably shorter and higher crania, with noticeably retracted and more upright nasal bones, and a more frontal orientation of the orbits (see the skulls in Fig. 1). Moreover, these late

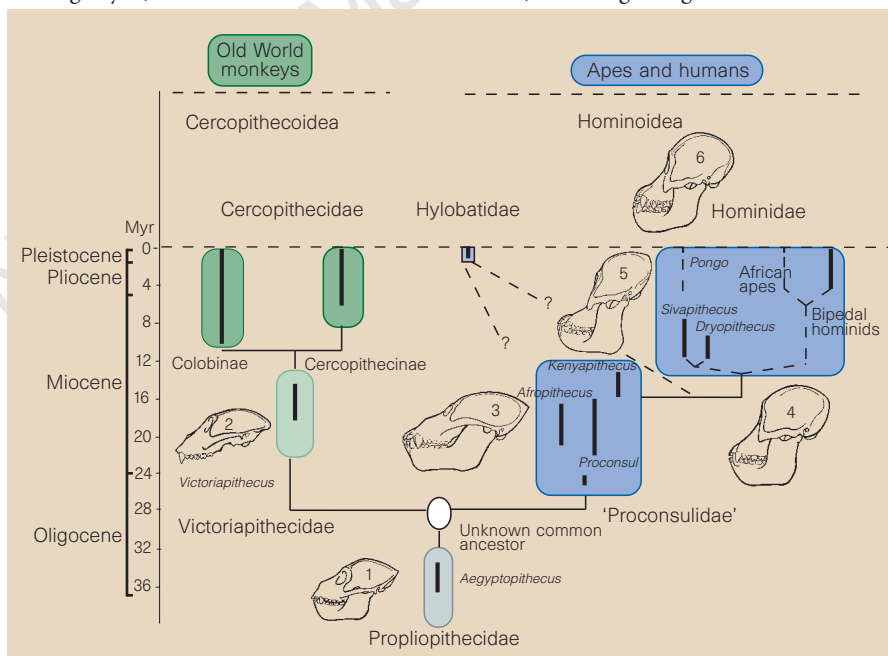


Figure 1 Simplified evolutionary tree of fossil and living Old World primates (catarrhines). 1, *Aegyptopithecus*; 2, *Victoriapithecus*, the specimen described by Benefit and McCrossin¹ (redrawn from their paper); 3, *Afropithecus*, neurocranium reconstructed; 4, *Dryopithecus*, reconstructed from Spanish and Hungarian fossils; 5, *Sivapithecus*, neurocranium reconstructed; 6: orang-utan, female. Note the similarities in overall skull shape between the stem catarrhines (Propliopithecidae, 1), the stem group of cercopithecids (Victoriapithecidae, 2) and the stem group of hominoids ('Proconsulidae', 3). Note also the differences between the ancestral African hominoids ('Proconsulidae', 3) and the group of the modern, late Miocene, Eurasian apes (4, 5) and living apes (6) (Hominidae).

Miocene forms completely match the modern great ape pattern, and this, in our opinion, makes them the only fossil hominoids that can be included in the family of great apes and humans.

There remain many divergent hypotheses about whether or not certain fossil hominoids are related to the living apes. What this diversity of opinion seems to indicate is that there is no consensus about which of the many different criteria and methods should be used for selecting characters and constructing theories. Analyses merely based on long lists of discrete characters, as demanded by cladistics⁸, do not seem to offer a way out of this impasse. We agree with Steve Ward⁹, who pointed out that our poor understanding of functional and developmental processes is a major hurdle that must be crossed before robust evolutionary hypotheses can be constructed.

Perhaps the time has come to reproach orthodox henningian cladistics⁸ for rejecting every kind of preselection of what Cuvier called "valuable" and "valueless" features. We would argue that a return to basics is

called for and that, before conducting any cladistic analysis, the usefulness of a character should be independently assessed in light of its underlying functional and structural constraints and developmental processes. Agreement on the central importance of this point may offer the first step towards forging a consensus about how to proceed. □

Meike Köhler and Salvador Moyà-Solà are in the Institut de Paleontologia M. Crusafont, c/Escola Industrial, 23, 08201 Sabadell, Barcelona, Spain.

1. Benefit, B. R. & McCrossin, M. L. *Nature* **388**, 368–371 (1997).
2. Andrews, P., Harrison, T., Delson, E., Bernor, R. & Martin, L. in *The Evolution of Western Eurasian Neogene Mammal Faunas* (eds Bernor, R. L., Fahlbusch, V. & Mittman, H.-W.) 168–207 (Columbia Univ. Press, New York, 1996).
3. Harrison, T. *J. Hum. Evol.* **16**, 41–80 (1987).
4. Fleagle, J. G. & Kay, R. in *New Interpretations of Ape and Human Ancestry* (eds Ciochon, R. L. & Corruccini, R. S.) 181–210 (Plenum, New York, 1983).
5. Starck, D. *Vergleichende Anatomie der Wirbeltiere* **2**, 3–776 (Springer, Berlin, 1979).
6. Hartwig, W. C. H. *Am. J. Phys. Anthropol.* **97**, 435–449 (1995).
7. Leakey, M. G., Leakey, R. E., Richtsmeier, J. T., Simons, E. & Walker, A. C. *Folia Primatol.* **56**, 65–85 (1991).
8. Henning, W. *Phylogenetische Systematik* (Parey, Hamburg, 1982).
9. Ward, S. in *Function, Phylogeny and Fossils* (eds Begun, D. R., Ward, C. V. & Rose, M. D.) 269–290 (Plenum, New York, 1997).



Figure 1 The issue of whether the banks of streams and rivers are better left to revert to forest or not is a complicated one, for streamside trees can either stabilize or erode the banks on which they grow. An interlocking network of roots can increase bank strength and therefore resist erosion. But trees that fall into streams can divert flow and trigger scour and local bank erosion where they cause the water flow to converge; where they cause flow to diverge, local deposition and aggradation of sediment results. In spite of local bank erosion, the increased variability of channel widths due to wood debris in forest streams contributes to the maintenance of a more variable and complex range of habitats.

for example, tree roots can substantially increase bank stability^{5–8} (Fig. 1). Clearing of the forest to the river banks along the Tolt river, a high-energy, gravel-bed channel in the Cascade range of Washington, triggered extensive channel widening because of the loss of bank-stabilizing tree roots. This widening increased the sediment supply downstream of the affected area, which in turn caused a pulse of gravel aggradation. At a finer scale, field surveys along an undisturbed reach of the Tolt river⁹ indicate that channel widening occurred only where log-jams diverted flow into banks. So scale also matters in assessing channel response; wood debris can cause local bank erosion within a channel reach in an area where the riparian forest overall maintains bank stability.

In some forest channels, logs and log-jams trap large volumes of sediment^{3,10,11} and offset sediment supplied from local bank scour. For example, channel-spanning log-jams on the Olympic peninsula, Washington state, can exceed 10 m in height, trap more than 10,000 m³ of sediment, and significantly lower reach-scale channel slopes¹¹. In these examples from channels with a mature riparian forest, the vertical storage of sediment impounded by organic debris outweighs storage lost from local channel widening. Log-jams also reduce rates of transport along the river bed by

River management

What's best on the banks?

David R. Montgomery

In the United States, strategies for the conservation of aquatic ecosystems emphasize the protection and restoration of streamside (riparian) forests^{1,2}—that is, current official policy encourages the reversion of streamside areas to mature forest. But it has not always been thus, and throughout most of the twentieth century US government agencies have actively removed wood debris from channels to improve navigation or allow fish easier passage. These practices began to change in the 1970s, after it was recognized that the accumulation of coarse organic debris from forests leads to the greater incidence of ecologically beneficial cover, slow-water refugia and pools, and increased sediment storage³.

Writing in *Geology*⁴, Stanley Trimble has introduced a new twist into this arena. He studied four reaches of Coon Creek, in the state of Wisconsin, each of which has grassed and forested stretches. His surveys show that channels flowing through forests tend to be wider than channels flowing through grasslands, an effect he attributes to the tendency for wood debris to concentrate flow into channel banks and erode them. From differences in average channel cross-sectional area, Trimble infers that clearing of riparian forests can decrease downstream sediment loads, and thereby help to decrease the downstream effects of land management, because grass cover can protect banks against erosion and increase local sediment deposition. Although he acknowledges the danger of arguing from a single case study,

Trimble warns that the restoration of riparian forests may not be good public policy.

These results and recommendations deserve close attention, as they run counter to the general understanding of the value and function of riparian forests. Indeed, another study⁵ published earlier this year came to the opposite conclusion: that bank erosion in forest channels accounts for a significantly smaller percentage of the sediment yield than in comparable grassland channels.

The trouble with this topic, as exemplified by these seemingly contradictory findings, is that prediction of the response of natural water-courses to man-made or natural change is a classic under-constrained problem. Many interrelated variables are involved, which moreover are subject to complex feedback and threshold responses. These variables include the interplay of sediment supply, size and lithology; the magnitude and frequency of water discharge; the nature of bank materials; the absence, or presence and type of, vegetation on banks; and the effect of flow obstructions such as wood debris. The type of channel and its history of disturbance, and the drainage-basin context, also affect channel response. This smorgasbord of influences means that simple guidelines and blanket generalizations rarely provide a sound basis for the management of rivers and streams.

Take the highly variable influence of bank materials on bank erosion⁶. In channels with banks composed of cohesionless materials,

decreasing energy gradients and increasing channel roughness^{12,13}. Even if forest channels are somewhat wider than grassland channels, they may still act as net sediment sinks.

Despite conflicting reports of the effect of trees on channel width, there does seem to be one consistent observation — that bankside forests increase the variance in channel form. Variable channel width translates into a greater range of flow depth, velocity and substrate size, resulting in increased variability or complexity of habitats; all of these are features that in my experience serve as a mantra for stream ecologists. Hence, evaluation of geomorphological and ecological impacts may yield contradictory conclusions; in some cases a little extra erosion may provide a more varied, diverse and productive habitat.

Trimble's paper⁴ illustrates the frustrating complexity facing those charged with managing, restoring and protecting the world's river systems. Single-recipe approaches provide a poor foundation for management of rivers and streams, in part because they often ignore connections between physical and biological processes. Moreover, coupling predictions of channel response to stream ecology and public-policy considerations further complicates an inherently challenging problem.

That leaves us with two distinct choices for ecologically orientated river management: either trust that 'natural is best' and promote restoration of riparian forests, or treat each river on a case-by-case basis. Trimble's cautionary thoughts on forest restoration would suggest the latter course of action. Unfortunately, few of those who work in this area receive the advanced training necessary to adequately understand and diagnose channel conditions. So, for now, perhaps it is wiser to favour nature's course. Continued research is needed into general models of channel response, and into the integration of fluvial geomorphology into river-system restoration and management. But the greater need is the establishment of a corps of river professionals whose expertise spans geomorphology, ecology and public policy. □

David R. Montgomery is in the Department of Geological Sciences, University of Washington, Seattle, Washington 98195, USA.

Protein folding

The difference with prokaryotes

Mary-Jane Gething

When proteins are made, they start off as chains of polypeptides. The study of how they come to fold into their correct three-dimensional structure is a wonderfully complicated business that has been tackled in numerous ways. In their paper on page 343 of this issue¹, Netzer and Hartl explore a somewhat neglected approach — that of looking at differences in the folding pattern of nascent polypeptides in the cytoplasm of prokaryotic and eukaryotic cells. In doing so, they tell us a great deal about some basic issues to do with protein folding, but we need to step back a little to understand why.

Anfinsen's classic *in vitro* experiments on the refolding of ribonuclease^{2,3} demonstrated that all of the information required to determine the final conformation of a protein can reside in the polypeptide chain itself: given the appropriate conditions, a denatured enzyme can refold into its native conformation in the absence of any other protein. Anfinsen's work was of seminal importance in establishing the central principle of self-assembly of proteins, but it probably delayed by a decade or two the initiation of investigations of protein folding *in vivo* — the trouble was we believed that the polypeptide sequence was all that there was to it. Nevertheless, over time we came to recognize that efficient folding *in vitro* is frequently limited to small, single-domain proteins and usually occurs under conditions of protein concentration, temperature, ionic strength and pH far removed from those occurring within cells.

The search for agents that support the folding of polypeptides under physiological conditions led to the characterization of a wide variety of molecular chaperones and folding catalysts that preside at the birth, maturation and death of most, if not all, proteins within cells^{4,5}. These chaperones and folding catalysts do not determine the final conformation of the polypeptide substrate (Anfinsen rules OK!); rather, they increase the efficiency of folding by inhibiting off-pathway aggregation reactions or catalyse rate-limiting isomerization steps. They are members of large protein families, and are present in all cell types, from prokaryotes to humans, and in all compartments of eukaryotic cells. The high degree of conservation of the members of each family during evolution has lulled us into believing that the process of protein folding is likely to be mechanistically similar in different organisms. This is where the new work of Netzer and Hartl¹ comes in.

The original premise underlying their study was that evolution of eukaryotic proteins by gene fusion or exon shuffling demands that the newly combined domains

must be able to fold independently of one another. Although this idea seems perfectly intuitive (at least in retrospect), to my knowledge its implications with respect to protein-folding mechanisms have not previously been tested. The study reveals that protein domains (which usually range from 100 to 300 amino acids in length) can indeed fold sequentially and independently when synthesized from a eukaryotic ribosome. In prokaryotes, however, folding of individual domains is delayed until the complete polypeptide chain has been synthesized. This latter situation can result in inappropriate interdomain interactions or aggregation — which are indeed often observed when complex eukaryotic proteins are synthesized in the prokaryote *Escherichia coli*.

Comparison of the protein-size profiles of different prokaryotic and eukaryotic organisms, made possible by the completion (or near completion) of sequence analyses of their genomes, reveals that prokaryotes may cope with this difficulty, at least in part, by restricting protein size and thus the complexity of protein-domain organization (see Fig. 1 of Netzer and Hartl's paper, page 343). Nevertheless, a lot of proteins in *E. coli* are more than 300 residues in length; many of them probably correspond to multifunctional proteins that have arisen by fusion of genes encoding the individual enzymatic activities or functions⁶. Given a post-translational mode of folding, such fusions would presumably have been productive only in cases where the individual domains could fold independently. In eukaryotes, however, the rapid evolution of multidomain proteins by gene fusion or exon shuffling (made possible following the development of complex genes) would not be restricted by any requirement that the component domains should not interfere with one another during the folding process.

Netzer and Hartl's work¹ also does much to resolve, intellectually if not mechanistically, a previously open question — why efficient protein folding in prokaryotes seems to rely so heavily and generally on the chaperonin GroEL, the eubacterial member of the Cpn60 (also known as Hsp60) family^{5,7}, whereas in the eukaryotic cytosol the chaperonin homologues (known as CCT^{7,8} or TRiC⁵) appear to play a much more limited role. GroEL and its co-chaperonin GroES form a cylindrical complex that encloses a central cavity in which polypeptides can fold sequestered from the cellular environment⁵. The observation that most polypeptides synthesized in *E. coli* can associate with GroEL prior to folding (F. U. Hartl, unpublished observations) may reflect the fact that they

1. National Research Council *Restoration of Aquatic Ecosystems* (National Academy Press, Washington DC, 1992).
2. National Research Council *Upstream: Salmon and Society in the Pacific Northwest* (National Academy Press, Washington DC, 1996).
3. Keller, E. A. & Swanson, F. J. *Earth Surf. Proc.* **4**, 361–380 (1979).
4. Trimble, S. W. *Geology* **25**, 467–469 (1997).
5. Stott, T. *Earth Surf. Proc.* **22**, 383–399 (1997).
6. Thorne, C. R. in *Vegetation and Erosion* (ed. Thornes, J. B.) 125–144 (Wiley, Chichester, 1990).
7. Shaler, N. S. *US Geol. Surv. 12th Annu. Rep. Pt I Geology* 219–345 (1891).
8. Gilbert, G. K. *US Geol. Surv. Prof. Pap.* 105 (Gov. Print. Office, Washington DC, 1917).
9. Montgomery, D. R., Buffington, J. M., Smith, R., Schmidt, K. M. & Pess, G. *Wat. Resour. Res.* **31**, 1097–1105 (1995).
10. Montgomery, D. R. *et al. Nature* **381**, 587–589 (1996).
11. Abbe, T. B. & Montgomery, D. R. *Abstracts with Programs – Geol. Soc. Am.* **28**(5), 41 (1996).
12. O'Connor, M. D. thesis, Univ. Washington, Seattle (1994).
13. Assani, A. A. & Petit, F. *Catena* **25**, 117–126 (1995).

must fold post-translationally. CCT chaperonins apparently interact with very few substrates, including actin and tubulin⁹. Netzer and Hartl point out that, although actin is a multidomain protein, its domains are composed of discontinuous amino-acid sequences, and therefore it would not be expected to fold co-translationally. So CCT chaperonins may have been retained in eukaryotes for the specific but limited task of coping with polypeptides that cannot fold in a stepwise fashion, domain by domain, as translation proceeds.

The molecular mechanisms underlying the different treatment of nascent polypeptides, following synthesis by prokaryotic and eukaryotic ribosomes, remain to be defined. Netzer and Hartl rule out the higher speed of translation in bacteria as a significant factor, but raise the possibilities that the prevention of folding of ribosome-bound polypeptides in *E. coli* might be effected either by ribosomal components or by molecular chaperones. One *E. coli* chaperone that can be exonerated is GroEL itself, because it does not recognize ribosome-bound chains⁵. By con-

trast, Hsp70 chaperones^{5,10} do interact with ribosome-associated nascent polypeptide chains, both in *E. coli* and in mammalian cells. Whether a previously unsuspected difference in the way in which prokaryotic and eukaryotic Hsp70 proteins interact with newly synthesized polypeptides underlies the failure of prokaryotes to support co-translational folding, or whether other translational co-factors are involved, is sure to be the subject of intensive studies in the coming months. □

Mary-Jane Gething is in the Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, Victoria 3052, Australia.

1. Netzer, W. J. & Hartl, F. U. *Nature* **388**, 343–349 (1997).
2. Anfinsen, C. B., Haber, E., Sela, M. & White, F. H. *Proc. Natl Acad. Sci. USA* **47**, 1309–1314 (1961).
3. Anfinsen, C. B. *Science* **181**, 223–230 (1973).
4. Gething, M.-J. & Sambrook, J. *Nature* **355**, 33–45 (1992).
5. Hartl, F. U. *Nature* **381**, 571–580 (1996).
6. Hawkins, A. R. & Lamb, H. K. *Eur. J. Biochem.* **232**, 7–18 (1995).
7. Horwich, A. L. & Willison, K. *Phil. Trans. R. Soc. Lond.* **339**, 313–326 (1993).
8. Kubota, H., Hynes, G. & Willison, K. *Eur. J. Biochem.* **230**, 3–16 (1995).
9. Lewis, S. A., Tian, G. L., Vainberg, I. E. & Cowan, N. J. *J. Cell Biol.* **132**, 1–4 (1996).
10. Rassow, J., von Ahsen, O., Bömer, U. & Pfanner, N. *Trends Cell Biol.* **7**, 129–133 (1997).

Extraterrestrial impacts

Earth rocked by combination punch

Dieter Stöffler and Philippe Claeys

The 100-km-diameter Popigai structure in Siberia is probably the fifth largest impact crater on Earth. Its age has been disputed — it could be anything from 5 to 65 million years (Myr) old. But this controversy appears to be finally settled: on page 365 of this issue¹, Bottomley *et al.* report ages of 35.7 ± 0.2 Myr for rocks from the Popigai crater that were melted by the impact. That means that 35.5 million years ago, in the space of only a few hundred thousand years — or perhaps much less — two large asteroids or comets struck the Earth, one in Siberia and another just off the west coast of North America.

The Late Eocene age for Popigai opens up intriguing questions about how two large impacts in quick succession might have affected the Earth. The period from the Late Eocene to the Eocene/Oligocene boundary (which is dated at 33.7 ± 0.5 Myr ago) is marked by the most severe mass extinction since the dinosaurs disappeared 65 Myr ago. The climate also worsened considerably in the Late Eocene, with cooling oceans and the first appearance of ice sheets in Antarctica.

Could these changes be the result of two almost coeval large impact events? We now know¹ that the Popigai crater is the same age as impact indicators (iridium and shocked quartz) found below the Eocene/Oligocene boundary in the Massignano section, Italy^{2–4}, which are dated radiometrically using volcanic ash layers at 35.7 ± 0.4 Myr

old^{2,3}. However, compositional evidence linking the Popigai crater and the Massignano ejecta is lacking; more work on this topic is urgently required. But the age coincidence is remarkable.

Excitement about the Popigai date grows in the light of the discovery a few years ago of a contemporaneous impact event on the other side of the world: the 85-km-diameter Chesapeake Bay impact structure, off the Virginia shore^{5,6}. The Chesapeake Bay crater had already been dated at 35.5 to 35.2 Myr old⁶. It is probably the source of the huge field of tektites (characteristic glassy nodules produced by melting of the target rock in an impact event) strewn mainly over Georgia, as indicated by the chemical similarities between the glass and the Chesapeake Bay target rocks^{7,8}. Similarly, the crater appears to be the source of microtektites and layers of excess iridium found on Barbados and in several ocean-floor drilling cores (such as DSDP 612), and dated at 35.4 ± 0.6 and 35.5 ± 0.3 Myr old^{9,10}, respectively.

So, within the error bars, the Popigai and Chesapeake Bay craters are the same age; at most, they occurred a few hundred thousand years apart. The tremendous energy released by these two successive huge impacts may have induced the long-lasting perturbation of the global environment which led to the terminal Eocene cataclysm and global changes.

We finally have a second solidly based case study, after the Cretaceous/Tertiary



100 YEARS AGO

We endeavoured before to ascertain what density of atmosphere might be conceded to the moon. This density being very feeble, it follows that the surface of our satellite must now be at a low temperature — at least, near the poles. There is even cause to ask if it is not totally or partially covered with ice. The most complete representation of the southern region ... inclines us to the opposite view — that is to say, that the presence of a great accumulation of ice must be considered improbable, as well for the polar caps as for the equatorial region. One is, therefore, led to imagine that the total moisture on the surface must have disappeared, undoubtedly by penetration into the interior of the globe, before the polar regions sank permanently beneath freezing-point. ... The cooling of our satellite, more rapidly than that of the earth, has shortened the period of condensation of the vapours. The water was filtered as fast as it was formed in innumerable volcanic orifices, which seemed prepared to receive it. Pl. vi. gives us an idea of the abundance of these openings in the neighbourhood of the pole, and one is led to think that the same constitution must have existed on the whole of the moon previously to the formation of seas.

From *Nature* 22 July 1897.

50 YEARS AGO

[Dr. Fosdick] does not deny that knowledge is dangerous, but he insists that the answer does not lie in trying to curb science or to fix boundaries beyond which intellectual adventure should not be allowed to go. The search for truth, he maintains, is, as it always has been, the noblest expression of the human spirit. Man's hunger for knowledge about himself, his environment and the forces by which he is surrounded, gives human life its meaning and purpose, and clothes it with final dignity. Civilisation is not being betrayed by science: the danger lies in the failure to use science wisely. ... The fundamental issue of our time is whether we can develop understanding and wisdom reliable enough to serve as a chart in working out the problems of human relations, or whether our lopsided progress will be allowed to develop to a point that capsizes our civilization in a catastrophe of immeasurable proportions. From *Nature* 26 July 1947.

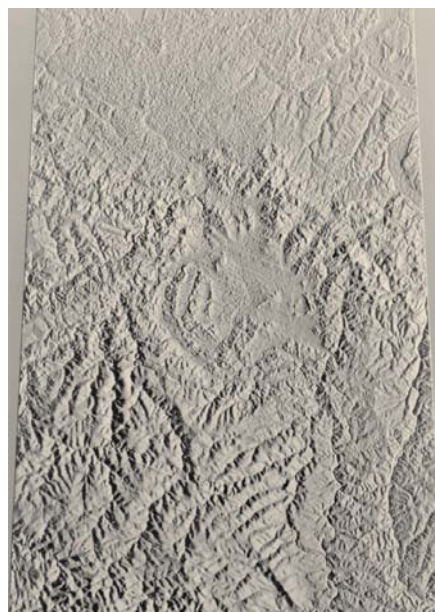


Figure 1 Popigai topography. The 100-km Popigai impact crater in Siberia shows up clearly on this topographic map. It has now been dated to around 35 million years ago, close to the time of another huge impact, and to a mass extinction in the Late Eocene.

(K/T) boundary, of the Earth's response to large impacts. In both cases, the location, size and age of the impact craters are known, there is a reasonable though not yet complete record of global or regional ejecta layers, and a detailed fossil succession is present in coeval, often continuous sediments from which a climatic and oceanographic record can be extracted.

In the case of the K/T boundary, 65 Myr ago, the coincidence and possible cause-effect relationship between the large impact which formed the 180–310-km-diameter Chicxulub crater in Yucatán, and the mass extinction of organisms, is striking. In the light of the K/T event, a rigorous assessment is needed of the fossil and climate record at small timescale across the Late Eocene. A concerted and open-minded effort¹¹ from palaeontologists, sedimentologists, geochemists and impact researchers should determine the role played by the Popigai and Chesapeake Bay impacts on the biological and climatic evolution of the Earth in the Late Eocene.

These two almost coeval impacts shed new light on the random behaviour of the terrestrial impact flux and the size–frequency distribution of impactors as a function of time. According to Bottomley *et al.*¹, “the expected repeat rate of impact structures of this size is of the order of ten million years”. So the question arises: how important are the statistics of small numbers for the effectiveness of impact-induced disturbances in the evolution of the Earth, and the related mass extinction events and climate changes in the past 600 million years? □

Dieter Stöffler and Philippe Claeys are at the Institut für Mineralogie, Museum für Naturkunde, Invalidenstrasse 43, D-10115 Berlin, Germany.
e-mail: dieter.stoeffler@rz.hu-berlin.de;
philippe.claeys@rz.hu-berlin.de

1. Bottomley, R., Grieve, R., York, D. & Masaitis, V. *Nature* **388**, 365–368 (1997).
2. Montanari, A. in *Global Catastrophes in Earth History: An Interdisciplinary Conference on Impacts, Volcanism, and Mass Mortality* (eds Sharpton, V. L. & Ward, P. D.) 607–616 (Geol. Soc. Am., Boulder, CO, 1990).
3. Montanari, A. *Eos* **71**, 1425 (1990).

4. Langenhorst, F. *Geology* **24**, 487–490 (1996).
5. Poag, C. W., Powars, D. S., Poppe, L. J. & Mixon, R. B. *Geology* **22**, 691–694 (1994).
6. Poag, C. W. & Aubry, M.-P. *Palaeos* **10**, 16–43 (1995).
7. Koeberl, C., Poag, C., Reimold, W. & Brandt, D. *Science* **271**, 1263–1266 (1996).
8. Shaw, H. F. & Wasserburg, G. J. *Earth Planet. Sci. Lett.* **60**, 155–177 (1982).
9. Obradovich, J. D., Snee, L. W. & Izett, G. A. *Abstracts with Programs – Geol. Soc. Am.* **21**, 134 (1989).
10. Glass, B. P., Hall, C. M. & York, D. *Chem. Geol.: Isotope Geosci. Sect.* **59**, 181–186 (1986).
11. Jablonski, D. *Nature* **387**, 354–355 (1997).

Homogeneous catalysis

On the trail of dioxygen activation

Craig L. Hill and Ira A. Weinstock

An ‘ideal’ homogeneous catalyst for dioxygen (O_2) oxidations would be selective, useful for tackling a variety of substrates and chemical structures, reasonably fast, and stable. Critically, and particular to the use of O_2 , the process should not require the wasteful addition of reducing agents. Such a system has remained a dream, but the work described by Neumann and Dahan on page 353 of this issue¹ constitutes notable progress towards realizing it.

All living matter and synthetic organic materials within the biosphere share a common condition and fate: they are thermodynamically unstable with respect to CO_2 and H_2O , products of reaction with O_2 . Fortunately for us, however, O_2 is a diradical, possessing a triplet ground-state electronic configuration (two unpaired electrons), whereas almost everything else possesses singlet ground states (paired electrons only). Because of this difference in ground electronic states, strong electronic interactions, which go hand-in-hand with low activation barriers to chemical reactivity, are spin-forbidden. This kinetic barrier is what makes the existence of life (as well as most synthetic organic materials) possible.

On the other hand, once dioxygen is ‘activated’ (that is, greater reactivity is induced), its reactions with organic substrates are difficult to control. This is because the diradical nature of dioxygen facilitates the formation of highly reactive and nonselective radical intermediates and radical-chain processes. Furthermore, because of the thermodynamic instability, such oxidations, once begun, are often highly exothermic and the resulting temperature increases further decrease selectivity. In consequence, the great majority of spontaneous O_2 reactions, such as those that lead to rancid butter, or the more rapid ones that occur during the combustion of fossil fuels, are nonselective and difficult to control.

Between the extremes of chemical inertia and uncontrollable radical-chain autoxidation processes lie opportunity and challenge. The opportunity: to make use of O_2 , the most abundant, inexpensive, energy-efficient and

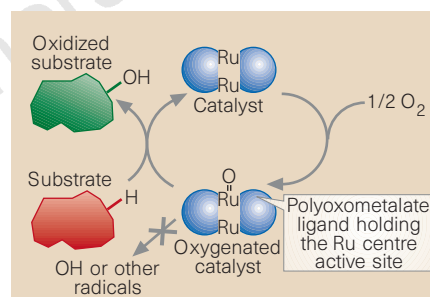


Figure 1 The Neumann–Dahan¹ catalyst, $\{[WZnRu_2(OH)(H_2O)](ZnW_9O_{34})_2\}^{11-}$, contains two sterically protected ruthenium atoms sandwiched between $[ZnW_9O_{34}]^{12-}$ halves (represented here as semicircles) of the oxidatively resistant polyoxometalate. The catalyst facilitates direct O_2 activation (right side of the figure) and substrate oxidation (left side) without generating nonselective radicals. Oxygen is incorporated directly into the substrate, the sole source of the electrons needed for the reductive activation of O_2 .

environmentally compatible oxidant available. The challenge: to control the reactivity of activated O_2 so that high chemical selectivities, and thus useful products, might be obtained. At the same time, the catalytic chemistry of oxygen and its inorganic and organic derivatives, while remarkably rich, is painfully complex and experimentally difficult. The design of homogeneous catalysts for both activation and selective use of O_2 is an extraordinarily tough task.

A third issue pertains to material and energetic efficiency. Dioxygen has four oxidizing electron equivalents. A common activation strategy, however, is to ‘waste’ some of them by partially reducing O_2 to peroxides or other reduced oxygen species that are more active and selective than O_2 (refs 2, 3). Many extremely useful homogeneous systems catalyse the selective reactions of reduced oxygen compounds⁴. Reductive activation of O_2 can also be accomplished by *in situ* introduction of an external supply of electrons (as with the effective reduction of O_2 by the coenzyme NADH in oxidative phosphorylation); this is a strategy exemplified by many

An artist of the floating world



This triptych, *Landscape with the Whirlpools at Awa*, is one of over 140 works by the Japanese master Hiroshige (1707–1858) that feature in an exhibition to celebrate the bicentenary of his birth. The painting is typical of

Hiroshige's work, inspired by the ruggedness and hostility of the region around his home in Edo (present-day Tokyo). An instinctive feeling for the violence of the natural world is also seen in his most famous series of

prints, *The 53 Stations of the Tokaido Road*, in which travellers battle through rain, mist and snow on the arduous journey from Edo to Kyoto. Hiroshige's work inspired many Western painters, particularly the French Impressionists. His prints are considered innovative for the time, and give some insight into a long-past Asian culture. The exhibition, at the Royal Academy of Arts in London, runs until 28 September 1997.

Kate McGown

highly selective biological catalyst systems.

However, in an ideal industrial catalyst system, the electron equivalents used for the reductive activation of O_2 would come from the target substrate itself. In a rare and much-cited example, Groves and Quinn demonstrated the epoxidation of alkenes by O_2 alone, catalysed by a ruthenium–porphyrin⁵. The hindered porphyrin ligand enabled key intermediates, $RuOORu$ and subsequently $Ru(IV)=O$ and $O=Ru(VI)=O$, to form, but not $RuORu$, the usual structural and kinetic dead-end species in homogeneous catalysed oxidations. The facility for these reactions precluded significant radical-chain oxidation. Ruthenium and other metal-oxo ($M=O$) species can participate in nonradical and highly selective reactions. Furthermore, the reactivities (rates and selectivities) of $M=O$ species are highly tunable by alteration of the other ligands on the metal centre, which has been accomplished in nature through evolution and in industry through human ingenuity.

An ideal catalyst must also be stable. Virtually all catalysts, from enzymes to synthetic metal complexes and heterogeneous metal surfaces, are subject to degradation during reaction. This leads to a loss in activity or selectivity, or both, often with catastrophic consequences. Homogeneous oxidation catalysts, which can include enzymes, ribozymes and many industrial catalysts, are particularly vulnerable as they contain organic ligands or structures that are thermodynamically unstable under operating (turnover) conditions. Like all organic compounds, the organic components of these catalysts are set up to fail; instability hangs over them like a sword of Damocles.

The generic oxidative instability problem of soluble oxidation catalysts has been solved by using particular derivatives of polyoxometalates (oxidatively resistant (d^0)

clusters of transition-metal cations and oxygen anions)^{6–9}. These derivatives are defect structures containing several oxygen atoms that simultaneously hold the metal ion (for example, Ru in Fig. 1), facilitating and controlling the metal ion's catalytic activity⁹. Neumann and Dahan¹ have used a sterically hindered Ru -containing polyoxometalate, $\{[WZnRu_2(OH)(H_2O)](ZnW_9O_{34})_2\}^{11-}$, that contains a barrier preventing the formation of dead-end MOM species (Fig. 1). This innovative system combines the direct activation of O_2 with subsequent selective substrate oxidation similar to the Groves–Quinn system, but with the stability of a polyoxometalate devoid of organic structure.

The precise nature of the oxygenated and active forms of the Neumann–Dahan catalyst and the mechanism of substrate oxidation are not fully elucidated, but this work is a significant step towards the Holy Grail of activating O_2 and catalysing selective non-radical oxidations without the consumption of other reagents. The challenges are now to achieve broader applicability (only hydrocarbons are addressed), faster reaction rates and compatibility with a range of reaction conditions. □

Craig L. Hill and Ira A. Weinstock are in the Department of Chemistry, Emory University, 1515 Pierce Drive, Atlanta, Georgia 30322, USA.

1. Neumann, R. & Dahan, M. *Nature* **388**, 353–355 (1997).
2. Sheldon, R. A. *J. Mol. Catal. A: Chem. (spec. iss.)* **117**, 1–489 (1997).
3. Barton, D. H. R., Martell, A. E. & Sawyer, D. T. (eds) *The Activation of Dioxygen and Homogeneous Catalytic Oxidation* (Plenum, New York, 1993).
4. Norrby, P. O., Kolb, H. C. & Sharpless, K. B. *J. Am. Chem. Soc.* **116**, 8470–8478 (1994).
5. Groves, J. T. & Quinn, R. *J. Am. Chem. Soc.* **107**, 5790–5792 (1985).
6. Pope, M. T. & Müller, A. *Angew. Chem. Int. Edn Engl.* **30**, 34–48 (1991).
7. Katsoulis, D. E. & Pope, M. T. *J. Am. Chem. Soc.* **106**, 2737–2738 (1984).
8. Hill, C. L. & Prosser-McCarthy, C. M. *Coord. Chem. Rev.* **143**, 407–455 (1995).
9. Hill, C. L. & Brown, R. B. *J. Am. Chem. Soc.* **108**, 536–538 (1986).

Daedalus

A personal thermostat

The human body is warmed by central heating. The furnace is the metabolizing organs such as the heart, liver and brain. The working fluid is the blood, pumped round the small-bore plumbing of the vascular system.

This arrangement works well in the tropics where it evolved; but in temperate and chilly climates it needs additional insulation (clothes), and even extra heating (fire). Even so, people with poor circulation often suffer from cold hands and feet. Daedalus has an upgrade.

He points out that blood contains both the fuel (glucose) and the oxidant (oxygen) for metabolic heating. In principle, it could generate heat by reaction wherever it is needed. In this connection he recalls an old scheme of his for countering alcoholism. A suspension of microspheres, coated with a set of immobilized alcohol-oxidizing enzymes, was injected into the subject's bloodstream. When he drank alcohol, they burnt it up as fast as it entered his blood. They replaced a pleasant intoxication by a sweaty fever.

Suppose, says Daedalus, the microspheres carried an enzyme system which oxidized not alcohol but glucose. By itself this would be a disaster. The body would burn all its fuel to heat, leaving none for anything else. But heat reversibly denatures some enzymes. It unwinds their protein chains a little, switching them off. So DREADCO biochemists are studying the oxidative biochemistry of lizards, molluscs, insects, and similar cold-blooded creatures. With luck, one of them will deploy a glucose-oxidizing enzyme which works well enough in the cold, but is reversibly denatured at 37°C — human body temperature. Immobilized on microspheres and injected into the bloodstream, the enzyme will be inert and harmless in the core regions of the body. But at the cold skin surface of an exposed extremity, it will renature. It will burn glucose until the blood around it has reached 37°C, when it will be switched off again. In effect, it will act as a perfect distributed body thermostat.

A DREADCO 'Warm Glow' injection should last for ages. The immobilized enzyme will be inaccessible to the immune system, and will degrade only slowly. The user will feel warm in the coldest conditions and the most provocative clothing. He or she will spend nothing on fuel, but a lot more on food — all of which will be burnt up in keeping warm. Even the greediest user will never get fat.

David Jones

Obituary

Jacques-Yves Cousteau (1910–97)

Inventor, film-maker and ocean explorer



Commandant Jacques-Yves Cousteau — a naval officer who became an explorer and film-maker — died of heart disease at his home in Paris on 25 June. His destiny is paradoxical. He is taken now for a scientist and a businessman, but in fact he was neither.

His reputation as a businessman arose from his talent for always finding the practical help and money needed to satisfy his all-embracing curiosity. His most famous invention, for example, came from the thought that it would be useful to have a way of being able to breathe while filming under water. He turned to the company Airliquide, where he met Emile Gagnan, an engineer who had developed a relief valve for feeding gas-fuelled engines. Together they developed a flexible relief-valve that enabled a diver to breathe compressed air from a gas bottle without choking. In 1943, following a filmed dive to a depth of 62 m, he filed a patent under the trade name Aqualung — the beginning of a great commercial success.

When, in 1950, Cousteau sought a ship with which to lead his own expeditions, he found a British benefactor, Sir Loel Guinness. With Guinness's help, he was able to buy an old mine-sweeper and convert it into an oceanographic vessel, the *Calypso*. Throughout his career he maintained the knack of obtaining funding for his costly expeditions, from *National Geographic*, the Centre National de la Recherche Scientifique, the oil companies BP and Elf, the industrial firms Fiat and Pechiney, and the American TV channels ABC and CNN.

Furthermore, Cousteau had the idea of setting up a kind of fan-club, so as to encourage donations of money from the

general public. First, in 1974, he established the Cousteau Society in America and then, in 1980, La Fondation Cousteau in France. Hundreds of thousands of members, seduced by Cousteau's fight for the exploration and defence of our planet, helped him, through their subscriptions, to finance his expeditions. Cousteau knew how to find the money he needed, but he wasn't a businessman seeking to make money for himself or those around him. All the proceeds from his ventures were reinvested in exploration and film production.

Cousteau wasn't a scientist either — his training as a naval officer made him more of an engineer than a scientist. But all his life, he rubbed shoulders with science as well as technology. To be able to film underwater, he had to continually invent: an autonomous diving system (the aqualung), waterproof cameras, underwater lighting, 'diving saucers' (able to dive and move about on the sea bottom), sea-scooters.... Knowing what he wanted, knowing whose doors to knock on, he always found the people and equipment he needed to realize his technical feats. In this way, he brought in a new era of underwater diving.

In organizing expeditions for scientists, he also provided a shot in the arm for oceanography, which had been advancing in fits and starts since the initiatives of Prince Albert the First of Monaco at the beginning of the century. By inviting a wide range of scientists — geologists, marine biologists, botanists, zoologists, archaeologists and biochemists — on board the *Calypso*, one might even venture that he invented marine field ecology, enabling a cross-fertilization of ideas, approaches and findings.

When, in 1966, the French government launched its own oceanographic vessel, the *Jean-Charcot*, Cousteau devoted himself exclusively to producing films for television. From that time on, the scientists invited on board *Calypso* were doing no more than fitting in with the needs of the film crews. When exploring the great rivers of the world (St Lawrence, Mississippi, Nile, Amazon, Danube), the researchers served, above all, as a guarantee to local authorities that no monkey business was afoot. And because representatives of national universities were on board, the authorities allowed access to inhospitable or forbidden regions. In exchange, Cousteau made available to the host country all the results of his team's research. Most important of all, for Cousteau, was to bring back films for the general public.

This shift from research to cinema

provoked the deepest distrust within the scientific community. All the more so because Cousteau had let it be written everywhere — including in his own publications — that he was a great scholar. Of course, one can be a great connoisseur of the oceans and sea floor without being the slightest bit scientific. But for Cousteau, science was not confined to the laboratory and the literature. To explore the unknown was, for him, the very essence of the scientific process. To discover a wreck, to explore an underwater cave, to encounter a new species of fish was, for him, a scientific work. Especially if, rather than keeping the discovery secret, it was broadcast to the world's television and cinema screens; and even if, in the haste of so doing, it meant making mistakes or approximations.

In his last work *L'homme, la Pieuvre et L'orchidée* (*Man, the Octopus and the Orchid*), published the day after his death, Cousteau admits that he had had only "brief encounters with science". But he can't find words harsh enough to denounce military and industrial secrecy in fundamental research. "All those who maintain secrecy over discoveries that might affect human health and life, they are nothing less than criminals", he writes, claiming to have published all his findings on the physiology of diving because he valued a diver's health above manufacturers' secrets.

This popularization is not much appreciated by the research fraternity. Their displeasure is greatest when the popularization is by way of mass-appeal television programmes, where one sees more of the man, Cousteau, and his divers, than the science. But Cousteau himself was emphatic about his approach, even if it later meant making U-turns, as he did in eventually criticizing the discharges of bauxite into the sea by Pechiney and the French nuclear tests at Mururoa, and in recognizing that the supposed 'death' of the Mediterranean Sea had failed to happen.

The fact remains that Cousteau, through his technological innovations, gave a healthy lift to oceanography after the Second World War. With his films, he has also aroused the curiosity of millions of television viewers, and shown, using pictures, how our planet is both rich and fragile. As for science with a big 'S', he says this in his posthumous work: "We have succeeded in using science to conquer nature. We could surely use it to conquer human nature." And, presumably, make humankind less destructive.

Roger Cans

Roger Cans, a former environmental correspondent for *Le Monde*, is at 6 rue de l'église, 72510 St-Jean-de-la-Motte, France.

Recent subsidence of the northern Suez canal

In contrast to a recent interpretation of delta coast stability¹, we now show that the northeastern Nile delta in Egypt has been actively sinking relative to the sea during the recent Holocene epoch. The northern Nile delta is only about 1 m above sea level, making the northern Suez Canal and coastal cities of Port Said and Port Fuad (combined population nearly 500,000) highly vulnerable. Subsidence and world sea-level rises contribute to coastal erosion, incursion of salt in the groundwater underlying the delta plain, and silting problems in the canal entrances². These processes must be considered when implementing protection measures for this area.

The entire Holocene sedimentary sequence has been recovered from Smithsonian core S-21 (total length, 49 m; elevation < 1 m), drilled in Port Fuad beside the northern canal bypass (Fig. 1). The early Holocene sediments comprise basal shell-rich sands identified as transgressive deposits from shallow marine to littoral origin. These are overlaid by a thick (47 m), typical Holocene deltaic offlap sequence (Fig. 1). Marine muds and delta-front sandy muds evolve upwards to sandy littoral and muddy lagoonal sediments.

Our study differs from earlier analyses^{3,4} in that we dated sediments using accelerator mass spectrometric (AMS) ¹⁴C analysis of individual shells, urchin fragments and wood, calibrating the dates to obtain calendar ages. We also took into account the water depth at the time of sediment deposition, estimated from an analysis of the fauna present (ostracod, foraminifera, mollusc). This is an important factor limiting the precision of subsidence-rate determination.

We obtained 11 AMS dates from core S-21 ranging from 6,480 ± 40 yr BP (radiocarbon dates are all calibrated calendar ages unless otherwise noted) at a depth of 45 m in the lower part of the Holocene mud section, to modern dates near the core top. There was a remarkably consistent pattern of sediment accumulation between 6,500 and 950 yr BP (at a depth of 14 m; Fig. 1). The upper 14 m of core section comprises a mixture of sandy shallow near-shore, strandline and coastal-ridge deposits — this inconsistency in the relationship between age and depth in the upper 14 m of sand is probably due to reworking during dredging of the Suez Canal bypass in the 1970s and 1980s.

Near the base of core S-21 (at a depth of 47 m), there is a marked unconformity (or break) between the top of the transgressive

sands and the immediately overlying marine prodelta muds. Sands at a depth of 48 m (1 m below the sand–mud contact) are dated to 8,540 ± 150 calendar yr BP by conventional ¹⁴C analysis of bulk-sample carbonate. This date is comparable to that of a sample near the top of transgressive sands in core Fadda-1 to the south (9,020 ± 330 calendar yr BP; uncalibrated age 8,480 ± 280 yr BP, at a depth of 42 m)⁵. Dates in both cores indicate a hiatus of up to 1,000 years.

Littoral shell-rich sands present at 48 m (~8,500 yr BP) and 14.6 m (~950 yr BP) were deposited at a water depth of 0–10 m, whereas the fauna in the muds at 45 m (6,480 yr BP) represents water depths of > 10 m (ref. 6). At 8,500 yr BP, sands were deposited at a depth of 0–10 m below a sea level which, taking a rise in eustatic sea level⁷ into account, was ~14 m below the present level; that is, 14–24 m below the sea

level at the time. At 6,500 yr BP, these sands were 3 m below the sediment surface, which was being deposited at a depth > 10 m at a time when sea level was ~6.5 m below present. Thus the 8,500-year-old sands would have sunk by that time to a depth of at least 19.5 m below present sea level.

Similar reasoning places these sands at a depth of 34.5 to 44.5 m below present sea level at 950 yr BP (0.5 m lower eustatic sea level + 34 m overlying sediments + 0–10 m water depth). At present, the sands lie at 48 m below sea level.

From this analysis we calculate a mean subsidence rate of 3.98 mm per year from 8,540 years ago to the present. Thick Holocene sediments in cores P.S.16-1979 (in Port Said)³ and Fadda-1 (ref. 5) at 46.5 m and 42 m, respectively, indicate similar subsidence rates. In addition to long-term subsidence, the absolute sea level is rising at roughly 1.0–1.5 mm per year⁸, so current relative sea-level rise in this area is at least 5 mm per year. Independent analysis of the 1923–1946 tide-gauge record at Port Said shows a relative sea-level rise of 4.8 mm per year⁹.

The accumulation rate of muds during most of the Holocene (5.4 mm yr⁻¹) is quite

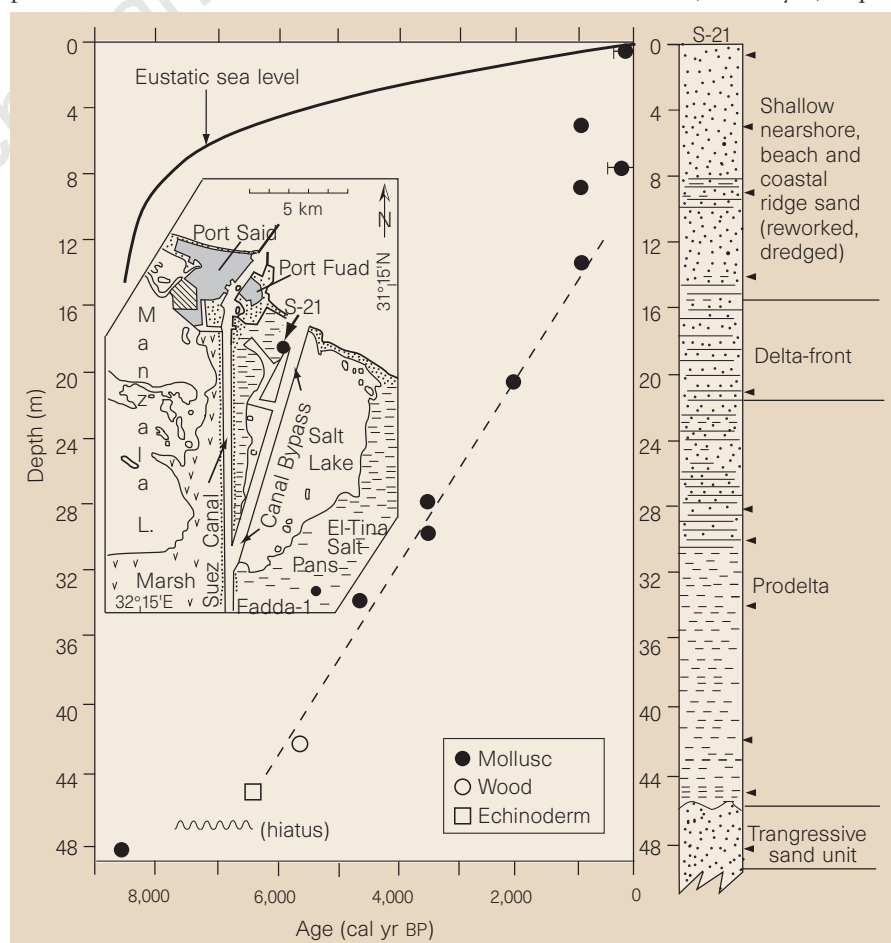


Figure 1 Calibrated radiocarbon dates (for $\Delta R=0$; that is, the worldwide value for the marine radiocarbon reservoir age, R , is assumed) from Nile delta core S-21, in relation to depth. Error bars (1σ) fall within the symbols except where shown. The dashed line shows the age–depth trend for 14–45 m depth, and the solid line is the eustatic sea-level curve⁷. Simplified core log (right) is from refs 3 and 6. Inset shows the core location in the northeastern Nile delta.

constant (dashed line, Fig. 1). After the deposition of littoral sands at 8,500 yr BP, inundation due to rapid sea-level rise allowed the accumulation of muds at a rate exceeding subsidence. Starting at 950 yr BP, more rapid accumulation of sands ($\sim 15 \text{ mm yr}^{-1}$) brought the sediment level up to sea level, and allowed deposition to keep pace with subsidence.

The main cause of subsidence in this delta region is ongoing faulting, as well as downwarping, of the underlying 3000 m of Late Miocene to Quaternary sequences¹⁰. The whole northern Nile delta plain has been lowered north of a flexure roughly 30 km inland from the coast². Vertical displacement of our dated Holocene sediments provides the most accurate means to measure long-term subsidence, rates of relative sea-level rise and recent sediment accumulation in this region.

Daniel J. Stanley

*Deltas-Global Change Program,
Smithsonian Institution,
Washington DC 20560, USA
e-mail: stanley.daniel@nmnh.si.edu*

Glenn A. Goodfriend

*Geophysical Laboratory,
Carnegie Institution of Washington,
Washington DC 20015, USA*

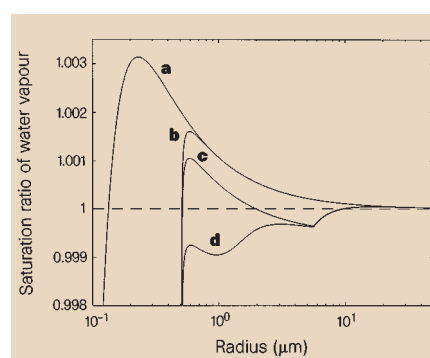
1. Said, R. *The River Nile: Geology, Hydrology and Utilization* (Pergamon, New York, 1993).
2. Stanley, D. J. & Warne, A. G. *Science* **260**, 628–634 (1993).
3. Stanley, D. J. *Science* **240**, 497–500 (1988).
4. Stanley, D. J. *Mar. Geol.* **94**, 147–154 (1990).
5. Sneh, A. *et al. Quat. Res.* **26**, 194–206 (1986).
6. Bernasconi, M. P., Stanley, D. J. & DiGeronimo, J. *Mar. Geol.* **99**, 29–43 (1991).
7. Lighty, R. G., Macintyre, I. G. & Stuckenrath, R. *Coral Reefs* **1**, 125–130 (1982).
8. Zerbini, S. *et al. Global Planet. Change* **14**, 1–48 (1996).
9. Emery, K. O., Aubrey, D. G. & Goldsmith, V. *Mar. Geol.* **81**, 41–52 (1988).
10. Neev, D. *et al. Israel Geol. Surv. Rep. MG/73/5* 1–43 (Israel Geol. Surv., Jerusalem, 1973).

Clouds without supersaturation

Traditional Köhler theory¹ describes the equilibrium vapour-pressure relationship between liquid solution particles and humid air. Here we present the concept of multiphase-multicomponent Köhler theory, which reveals that stable cloud droplets of size 1–10 μm could exist in air with a relative humidity of less than 100%. This may explain the occurrence of persistent large-droplet fogs or smogs such as previously existed in London and which are now found in various heavily polluted locations, near the exits of chimneys and in the plumes of volcanoes.

Assumptions are usually made, explicitly or implicitly, that the solute mass in a droplet is fixed as it grows in size. The result is that a critical size exists at which the

Figure 1a, Conventional Köhler curve for a 30 nm dry particle consisting of ammonium sulphate at 298 K. **b**, A similar particle containing an additional, insoluble 500 nm core (the amount of ammonium sulphate is the same as with a). The effect of insoluble material has been studied in detail⁷. **c**, Particle containing a slightly soluble 500 nm core. The solubility used⁸ ($0.00209 \text{ g cm}^{-3}$) corresponds to that of CaSO_4 . The sharp minimum of the curve shows the point at which all of the core is dissolved. CaSO_4 particles occur commonly in air, and dissolved CaSO_4 has been found in fog-water collected in the Po Valley, Italy⁹. **d**, Effect of an added, highly soluble gas, nitric acid. Initial gas-phase concentration of HNO_3 was 1 p.p.b.v., and the Henry's law constant used¹⁰ (mole-fraction scale) was 853.1 atm^{-1} . Because nitric acid is allowed to deplete from the gas as it is absorbed by the droplets, the term $b_a(r)$ in equation (1) depends on the aerosol number concentration, which in this case was assumed to be 100 cm^{-3} . Aerosol size distribution was taken to be monodisperse (a qualitatively similar curve would result if the aerosol population was $1,000 \text{ cm}^{-3}$ and initial HNO_3 concentration 3 p.p.b.v.). The smooth minimum in the curve is caused by the depletion.



Raoult effect (depression of vapour pressure due to dissolved substance) just balances the Kelvin effect (increase of vapour pressure due to droplet curvature). Above that size, a droplet is said to be activated and will grow spontaneously if the ambient supersaturation remains at or above the respective equilibrium value. However, this ordinary Köhler formulation does not describe droplets in which the solute derives from either slightly soluble aerosol particles² or soluble gases^{3,4}.

We consider a situation in which aerosol particles containing both hygroscopic matter (salt) and a weakly soluble or insoluble core, are suspended in humid, polluted air. The particles absorb water, highly soluble gases such as nitric acid (which are depleted from the vapour due to the uptake by the particles), and weakly soluble gases (with a roughly constant gas concentration). The weakly soluble core is assumed to keep the surrounding liquid as a saturated solution until it has fully dissolved. At equilibrium, the saturation ratio of water vapour (S) equals the equilibrium vapour pressure ($p_{\text{sol,w}}$) of water over a solution droplet of a given radius (r) divided by the water vapour pressure over a plane surface ($p_{\text{s,w}}$), which is given by:

$$S = \frac{p_{\text{sol,w}}}{p_{\text{s,w}}} \quad (1)$$

$$\approx 1 + \frac{a}{r} - x_{\text{ws}} - x_{\text{wg}} - \frac{b_s}{(r^3 - r_0^3)} - \frac{b_a(r)}{(r^3 - r_0^3)}$$

Here r_0 is the radius of the core, a/r describes the Kelvin effect, x_{ws} and x_{wg} are the mole fractions of the weakly soluble species originating in the aerosol particle and in the gas phase, respectively, $b_s/(r^3 - r_0^3)$ describes the Raoult effect of the soluble salt, and $b_a(r)$ is the term accounting for the highly soluble, condensing trace gas⁵.

Figure 1 compares the traditional Köhler curve to the more complete multiphase-multicomponent theory. The particle sizes

and compositions are selected as realistic examples as are the concentrations of water-soluble gases. We have not included in the curves the effect of a slightly soluble gas. For simplicity, we assume that the highly soluble substances do not influence the solubility of the weakly soluble substances (a conservative approximation). What becomes evident is that the combination of the amount of solute from the soluble gas and/or the slightly soluble aerosol particle, along with the vapour pressure reduction due to the Kelvin effect for the increased initial size of the slightly soluble particle, may be sufficient to keep the saturation ratio below unity for droplet sizes up to $\sim 10 \mu\text{m}$. In retrospect, this result is evident in the traditional simple formulation of the Köhler equation⁶:

$$S = 1 + \frac{a}{r} - \frac{b}{r^3} \quad (2)$$

because $S < 1$ if $b/r^3 > a/r$, which simply depends on the magnitudes of a and b . What we have pointed out is that realistic particle sizes, particle solubilities and gas concentrations exist to cause this to happen. In fogs formed by these processes, the cloud-drop-sized particles are thermodynamically similar to submicrometre 'haze' particles. However, without sophisticated physicochemical methods, they are physically indistinguishable from classical, 'activated' drops.

Markku Kulmala, Ari Laaksonen

*Department of Physics, PO Box 9,
FIN-00014, University of Helsinki, Finland
e-mail: kulmala@phc.u.helsinki.fi*

Robert J. Charlson

*Departments of Atmospheric Sciences
and Chemistry,
University of Washington,
Box 351700 Seattle, Washington 98195, USA*

Pekka Korhonen

*Finnish Meteorological Institute,
Air Quality Department, Sahaajankatu 20 E,
FIN-00810, Helsinki, Finland*

1. Köhler, H. *Trans. Far. Soc.* **32**, 1152–1161 (1936).
2. Shulman, M. L., Jacobson, M. C., Charlson, R. J., Synovec, R. E. & Young, T. E. *Geophys. Res. Lett.* **23**, 277–280 (1996).
3. Kulmala, M. *et al.* *J. Geophys. Res.* **98**, 22949–22958 (1993).
4. Korhonen, P., Kulmala, M. & Vesala, T. *Atmos. Env.* **30**, 1773–1785 (1996).
5. Laaksonen, A., Korhonen, P., Kulmala, M. & Charlson, R. J. *J. Atmos. Sci.* (submitted).
6. Pruppacher, H. R. & Klett, J. D. *Microphysics of Clouds and Precipitation* 142 (Reidel, Dordrecht, 1980).
7. Hänel G. *Adv. Geophys.* **19**, 73–188 (1976).
8. *CRC Handbook of Chemistry and Physics* 74th edn (eds Lide, D. R. & Frederikse H. P. R.) 4–49 (CRC, Boca Raton, 1993).
9. Del Monte, M. & Rossi, P. *Atmos. Environ.* **31**, 1637–1646 (1997).
10. Brimblecombe, P. & Clegg, S. L. *Atmos. Env.* **24A**, 1945–1955 (1990).

A new west African chimpanzee subspecies?

We have sequenced a portion of the mitochondrial DNA control region from the hairs of Nigerian chimpanzees living on both sides of the Niger River, which is proposed as a biogeographic boundary. Our data suggest that a previously unrecognized type of chimpanzee may be present in Nigeria and adjacent parts of Cameroon, and that the zoogeographic barriers separating chimpanzees are different from those previously proposed.

Three subspecies of the common chimpanzee are generally recognized: west African *Pan troglodytes verus*, central African *P. t. troglodytes* and east African *P. t. schweinfurthii* with the lower Niger River separating *P. t. verus* and *P. t. troglodytes*¹. Mitochondrial DNA (mtDNA) sequences² are consistent with this view although they tentatively elevate *P. t. verus* to the status of a full species (*Pan verus*). Given the 1,700-km gap in the sampling distribution between western and central African chimpanzees in ref. 2 (Fig. 1a), the long-term genetic isolation of *P. t. verus* populations has been questioned³.

To investigate geographically intermediate populations and the phylogeographic 'break' between subspecies believed to occur at the lower Niger, we studied a 600-km span

of Nigeria (Fig. 1b), collecting 12 chimpanzee hair samples from nests, captive individuals and skins. From those samples, we sequenced 414 base pairs of the mtDNA control region (D-loop). A combined data set of 87 individuals, including sequences obtained from GenBank⁴, was analysed using the neighbour-joining method and maximum parsimony and likelihood searches⁵, to give a 50% majority-rule consensus tree based on 10,400 equally parsimonious trees (Fig. 2).

All the Nigerian chimpanzees, from both sides of the Niger, form a single, distinct cluster, most closely allied with the *P. t. verus* clade. Two genetically distinct chimpanzee forms are inferred in west Africa: one group restricted to far-west Africa (Senegal to Ghana) and the other from Nigeria and probably adjacent parts of Cameroon.

In the sequenced mtDNA region, the two west African lineages differ at least as much as do the eastern and central forms. These differences might be the result of two alternative taxonomies. Either Nigerian chimpanzees are classified within *P. t. verus*, and *P. t. troglodytes* and *P. t. schweinfurthii* are collapsed into another subspecies, or Nigerian chimpanzees are a separate subspecies, along with the other three named subspecies. If a fourth subspecies is eventually recognized, the name *vellerosus* seems to be available⁶. These proposals need to be confirmed by more extensive sampling, sequencing of several independent nuclear loci and morphological data.

The Niger River has been proposed to separate western and central chimpanzees. Our findings suggest, however, that the Niger has not been a substantial boundary to gene exchange. Instead, chimpanzees in eastern Nigeria and those south of the Sanaga River in Cameroon, separated by only 200 km, have sufficiently divergent sequences to place them in the western and central chimpanzee clades, respectively. We therefore suggest that a phylogeographic discontinuity between western and central chimpanzees is located somewhere between the Cameroon Highlands in Nigeria and the banks of the Sanaga River in

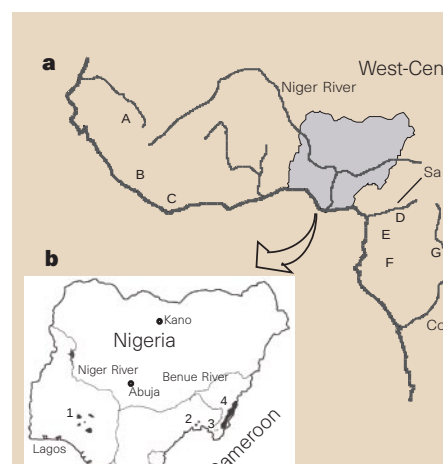


Figure 2 Phylogeny derived using a maximum parsimony analysis of chimpanzee mtDNA D-loop sequences. Triangles at the tips of branches represent a monophyletic cluster of individuals. Letters or numbers after names refer to localities in Fig. 1.

Cameroon. Distribution data for several other primates suggest the Sanaga River is an important boundary for many taxa⁷. A major genetic discontinuity has also been noted for gorillas from western and eastern Africa^{8,9}.

West African chimpanzees are the most threatened with extinction¹⁰. Chimpanzees in Nigeria and western Cameroon are especially endangered, as a result of hunting and habitat destruction. These animals now appear to constitute a genetically distinct evolutionary unit which has been neglected in conservation planning and should be given immediate attention.

M. Katherine Gonder

John F. Oates

Hunter College and the Graduate School of the City University of New York, Department of Anthropology, Hunter College, 695 Park Avenue, New York, New York 10021, USA

Todd R. Disotell

Department of Anthropology, New York University, 25 Waverly Place, New York, New York 10003, USA

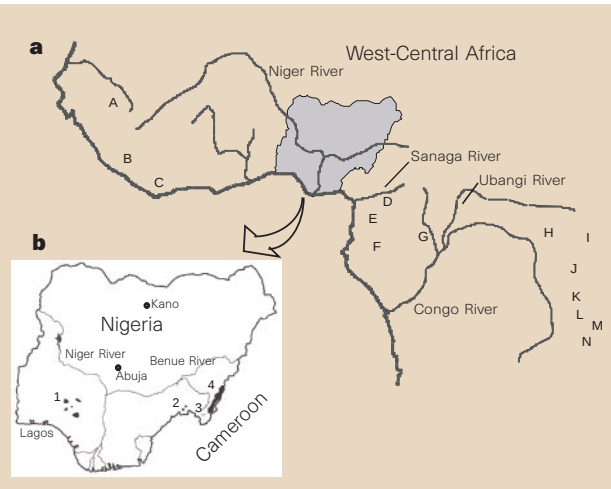
Michael R. J. Forstner

Juan Carlos Morales, Don J. Melnick*

*Center for Environmental Research and Conservation, and *Departments of Anthropology and Biological Sciences,*

Columbia University, 1200 Amsterdam Avenue, New York, New York 10027, USA

Figure 1a, Localities from which chimpanzee mtDNA was sampled by Morin *et al.*² **b**, Sources of new samples from Nigeria. 1, Ondo State forests (confiscated skins); 2, Donga Valley (captive); 3, Mambilla Plateau (nest hairs); 4, Gashaka-Gumti National Park (nest hairs).



1. Hill, W. C. O. in *The Chimpanzee* Vol. 1 (ed. Bourne, G. H.) 22–49 (Karger, Basel, 1969).
2. Morin, P. A. *et al.* *Science* **265**, 1193–1201 (1994).
3. Jolly, C., Oates J. F. & Disotell, T. R. *Science* **268**, 185–186 (1995).
4. Benson, D., Bojowski, M., Lipman, D. J. & Ostell, J. J. *Nucleic Acids Res.* **24**, 1–5 (1996).
5. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. in *Molecular Systematics* 2nd edn (eds Hillis, D. M., Moritz, C. & Mable, B. K.) 407–514 (Sinauer, Sunderland, MA, 1996).
6. Gray, J. E. *Proc. Zool. Soc. Lond.* 180–181 (1862).
7. Grubb, P. in *Vertebrates in the Tropics* (eds Peters, G. & Hutterer, R.) 187–214 (Museum Alexander Koenig, Bonn, 1990).
8. Ruvofo, M. *et al.* *Proc. Natl Acad. Sci. USA* **91**, 8900–8904 (1994).
9. Garner, K. J. & Ryder, O. A. *Mol. Phylogenet. Evol.* **6**, 39–48 (1996).
10. Teleki, G. in *Understanding Chimpanzees* (eds Heltne, P. G. & Marquardt, L. A.) 312–353 (Harvard Univ. Press, Cambridge, MA, 1989).

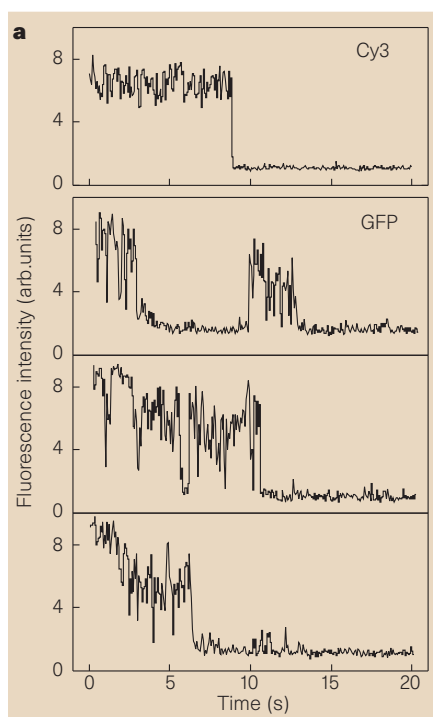
Imaging individual green fluorescent proteins

Recent advances in fluorescence microscopy techniques have allowed the video-time imaging of single molecules of fluorescent dyes covalently bound to proteins in aqueous environments¹. However, the techniques have not been exploited fully because proteins can be difficult to label, and dye modification may cause partial or complete loss of activity. These difficulties could be circumvented by fusing proteins to green fluorescent protein (GFP) of the jellyfish *Aequorea victoria*. Here we report that single S65T mutant GFP molecules² can be imaged using total internal reflection microscopy, and that ATP-driven movement of an individual kinesin molecule (a microtubule motor protein) fused to GFP can be readily observed.

We visualized bacterially expressed GFP or kinesin-GFP fusion proteins adsorbed at a low density to a quartz slide using a total internal reflection (TIR) fluorescence microscope that we had constructed³. Circularly polarized light at a wavelength of 488 nm from an argon laser was focused through a prism to produce total internal reflection and evanescent field illumination of an area roughly $30 \times 40 \mu\text{m}$ on the slide¹. Fluorescence was collected by an objective lens (Nikon PlanApo 100/1.40), collimated, passed through a custom-designed dichroic mirror and barrier filter (Chroma) and focused onto a charge-coupled device (CCD) camera with a selected intensifier tube (SR UB Gen3+, Stanford Photonics).

There are several reasons to conclude that single GFP molecules are imaged. First, spots disappeared in single quantal events, and not infrequently reappeared at a similar intensity, indicative of reversible photochemistry of a single molecule (Fig. 1). Second, the ratio of intensities of GFP spots (44.1 ± 1.3 arbitrary units, mean $\pm$ s.e.m.; $N=182$) relative to spots of the dye Cy3 (51.0 ± 1.6 ; $N=166$) is 0.86, as compared to a predicted ratio of 0.90 based on their extinction coefficients at 488 and 514 nm and the characteristics of our filter sets and camera. Cy3 spots, which show clear single-step photobleaching reactions (Fig. 1), have been shown^{1,4} to correspond to single molecules. Third, the fluorescent spot intensities of two fusion products of GFP and motor proteins (kinesin⁵ and Ncd⁶) that were shown to form homodimers by hydrodynamic methods were 1.9 and 2.1 times greater than spots containing GFP fused to a monomeric motor-protein construct (unc104)⁷.

Single GFP molecules showed greater



intensity fluctuations than the organic fluorophore Cy3, and their intensities generally declined during illumination (Fig. 1). This suggests that the protein-embedded chromophore has spectral dynamics that are distinct from organic dyes in solution. Light-driven spectral fluctuations have been observed⁸ for single dye molecules in polymethylmethacrylate films. As dynamics in the core of a protein⁹ may be analogous to those in glassy polymers¹⁰, we tentatively favour light-driven spectral fluctuations for the origin of the GFP intensity fluctuations.

The disappearance followed by re-appearance of fluorescence (Fig. 1) indicates that the chromophore undergoes a reversible reaction. The rate of photobleaching of GFP (0.12 s^{-1} at 10 mW) is faster than that of Cy3 (0.06 s^{-1} at 5 mW) and is not influenced by removal of molecular oxygen from the solution. The photobleaching rate of GFP was linearly related to laser power over the available range (2.5–20 mW).

To test the utility of GFP fusion proteins for measuring single molecule dynamics, we fused GFP to residue 560 of the human kinesin heavy chain. We purified the fusion product to homogeneity from an *Escherichia coli* expression system, and combined it with 1 mM ATP and sperm flagellar axonemes. As described previously for Cy3-labelled kinesin⁴, individual fluorescent spots bound to and moved processively along the axonemal microtubule substrate (Fig. 1). The fluorescence intensities of moving spots indicated that they are single molecules and not aggregates. We also found that kinesin-GFP exhibited

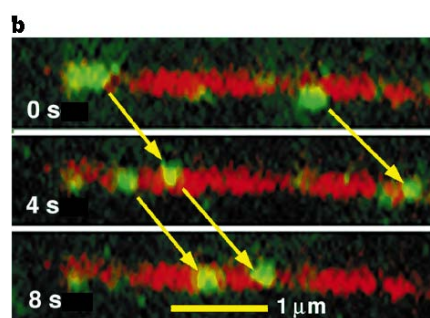


Figure 1a, Single-molecule intensity records for Cy3 dye and GFP (S65T mutant). Typical traces show the intensities of fluorescent spots (7×7 -pixel area in an 8-bit image, scaled by 1,000) for a kinesin-bound Cy3 dye⁴ (10 mW, 514 nm illumination) and GFP (20 mW, 488 nm illumination). Fluorescence from GFP disappears and reappears abruptly. There is a higher noise level (at comparable signal level) of GFP compared with Cy3, and the GFP fluorescence intensity declines after illumination. **b**, Individual kinesin-GFP molecules (pseudo-coloured green) moving along Cy5-labelled axonemal microtubules (pseudo-coloured red) in the presence of 1 mM ATP and previously described assay conditions⁴.

microtubule-based motility when expressed by *in vitro* transcription and translation in rabbit reticulocyte lysate or wheat-germ extract, and diluted into an ATP-containing buffer.

These GFP assays have advantages over our previously published Cy3 assay⁴ in ease as well as having a stoichiometric amount of fluorophore at a known position. We have extended our method to other motors that are difficult to purify and label by chemical modification.

The ability to measure kinesin-GFP molecule dynamics in real time also suggests broader applications of the technique. It may be possible to study single molecule dynamics of many types of proteins or protein complexes, particularly ones that are difficult to label by chemical means. Our preliminary data indicate that single molecules of GFP fused to membrane or juxta-membrane proteins can be imaged in living cells.

Daniel W. Pierce

Nora Hom-Booher

Ronald D. Vale

Howard Hughes Medical Institute
and Department of Pharmacology,
University of California, San Francisco,
California 94143-0450, USA
e-mail: vale@phy.ucsf.edu

- Funatsu, T., Harada, Y., Tokunaga, M., Saito, K. & Yanagida, Y. *Nature* **374**, 555–559 (1995).
- Heim, R., Cubitt, A. B. & Tsien, R. Y. *Nature* **373**, 663–664 (1995).
- Pierce, D. W. & Vale, R. D. *Methods Cell Biol.* (in the press).
- Vale, R. D. *et al.* *Nature* **380**, 451–453 (1996).
- Navone, F. *et al.* *J. Cell Biol.* **117**, 1263–1275 (1992).
- McDonald, H. B. & Goldstein, L. S. *Cell* **61**, 991–1000 (1990).
- Otsuka, A. J. *et al.* *Neuron* **6**, 113–122 (1991).
- Lu, H. P. & Xie, X. S. *Nature* **385**, 143–146 (1997).
- Ormo, M. *et al.* *Science* **273**, 1392–1395 (1996).
- Pierce, D. W. & Boxer, S. G. *J. Phys. Chem.* **96**, 5560–5566 (1992).

On the trail of the transistor

Crystal Fire: The Birth of the Information Age

by Michael Riordan and Lillian Hoddeson
Norton: 1997. Pp. 352. \$27.50

Frederick Seitz

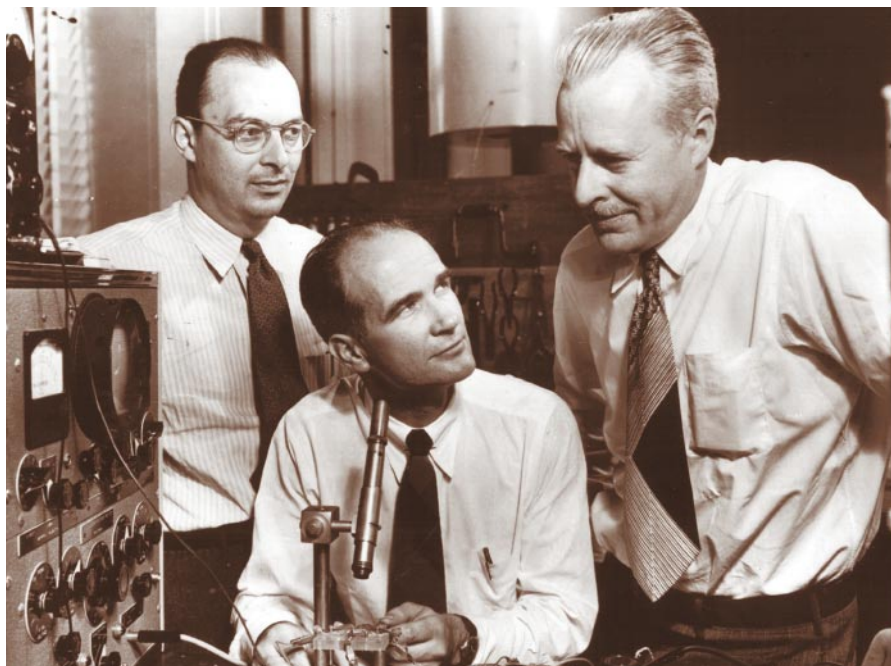
For a quarter of a century, Lillian Hoddeson, a physicist, has collected background material on the history of solid-state physics, both basic and applied, carrying out taped interviews with many of the most creatively active scientists, and studying the available literature, including archival files. One of the many by-products of this endeavour has been the multi-authored book *Out of the Crystal Maze* (Oxford University Press, 1992), of which Hoddeson was co-editor and leading catalyst.

Early in this decade, and in recognition of the fact that 1997 would be the fiftieth anniversary of the invention of the transistor, she joined with Michael Riordan, another physicist with a comparable interest in science history, to produce this lively account of the invention, with much emphasis on events that preceded and followed it. The background, personalities and actions of the principal actors are given ample coverage.

Mervin Kelly, who became head of the Bell Telephone Laboratories in 1936 and remained its leader for more than 20 years, can clearly be said to be the spiritual father of the transistor. While head of the vacuum-tube division of the Western Electric Company, also a subsidiary of AT&T, he noted in the late 1920s the remarkable, durable properties of inexpensive copper-oxide rectifiers, commonly used at that time for the conversion of relatively low-frequency alternating current to direct current. Among many virtues, they needed no heated filament to provide conduction electrons. He decided that the company should see if it could develop a comparable semiconductor triode.

Others might similarly ponder this matter, but Kelly would eventually gather together the appropriate combination of resources to achieve success. Once he became head of the laboratories, he took steps to increase the size of the group involved in solid-state research, which already contained Walter Brattain and Gerald Pearson. William Shockley was one of the first additions.

Unfortunately, Kelly's ambitious dream had to be all but shelved until 1945 because of the more immediate pressures of war research. In the meantime, however, much attention was devoted to the relative perfection of silicon and germanium diodes for use as 'mixers' in microwave radar — a development started in England and carried on in several laboratories in the United States, including the Bell Laboratories. Two new, versatile and more



John Bardeen, William Shockley and Walter Brattain (left to right) at Bell Telephone Laboratories in 1948. The picture appears in *The Invention that Changed the World: The Story of Radar from War to Peace* by Robert Buder, reviewed by Frederick Seitz in *Nature* 384, 424 (1996) and now published in the United Kingdom by Little, Brown at £20.

thoroughly understood semiconductors had become available for exploratory research.

With the end of the war, Kelly appointed Shockley as head of a team attempting to develop a triode. Fortunately, Shockley succeeded at this time in adding John Bardeen to the group. Bardeen had taken leave of an academic post in physics for military research early in the war and found the new programme in solid-state physics particularly challenging. He possessed two very valuable gifts: he had the tenacity of a bulldog when attacking a difficult problem and he possessed a remarkable 'third eye' which provided him with special insight into complex problems. His advent proved to be critical for the immediate course of the programme.

Early on, Shockley had tried to develop what is now known as a field-effect transistor in the most direct way, namely by attempting to alter the number of conducting carriers (electrons or holes) in a strip of semiconductor carrying a current by applying an external field normal to the strip, much as one might alter the charge in a condenser plate. The experiment failed.

Shockley, temporarily discouraged, turned the problem over to Bardeen and Brattain, who had developed a working partnership, while he turned to other solid-state problems. This was in fact one of Shockley's most creatively productive periods in relation to more general aspects of solid-state physics.

Bardeen and Brattain decided to try to

understand the factors that had caused Shockley's experiment to fail. Bardeen guessed that the specimen of semiconductor possessed surface trapping levels for the conducting carriers and that the occupation of these could vary in such a way as to compensate for variations in external applied field. With the aid of colleagues, they were able to prove that this was the case, opening the way to the eventual development of field-effect transistors.

Then, in re-examining a different failed experiment carried out in connection with the study of trapping levels, they made a remarkable discovery. When a current of minority carriers was injected from a point-contact electrode and induced to flow to a nearby point-collecting electrode, both that current and one of majority carriers from the collector to a third, more distant electrode (the base) could be modulated by altering the potential of the base. The invention of the bipolar point-contact transistor followed. It was demonstrated to management on 23 December 1947.

Because the invention was founded as much on empirical observation as on theoretical understanding, several issues about the operation of the device had to be explored. Among the most important, John Shive soon demonstrated that minority carriers actually can travel considerable distances through bulk material where majority carriers dominate. The minority current was not confined to a

thin depletion area at the surface, as Bardeen thought might be the case.

Shive's demonstration caused Shockley to propose an important modification of the original transistor. Instead of injecting and collecting minority carriers with the use of point contacts, one could employ relatively large area p-n junctions and gain the advantage of much larger currents. The concept of the bipolar junction transistor, which was to dominate the field for the next two decades or so, was born. A semiconductor triode was not only feasible but was about to generate one of the greatest technological revolutions in history.

The text is rich in anecdote, presenting the reader with colourful views of many of the secondary as well as the primary figures involved. Backgrounds, hopes and destinies are examined. Matters such as Shockley's gradual change in personality as he became a celebrity, and the changes in the working atmosphere which made Bardeen decide to return to an academic career, are dealt with in a forthright manner.

Although the book is written in a semi-popular style which should appeal to the type of general reader interested in the history of technology, the historical research on which it is based is impeccably sound, unlike many books of this kind. The authors, with their professional reputations involved, have left no stone unturned in making their work an enduring classic. □

Frederick Seitz is at the Rockefeller University, 1230 York Avenue, New York, New York 10021-6399, USA.

Ceaseless fluctuations

Conceptual Developments of Twentieth Century Field Theories

by Tian Yu Cao

Cambridge University Press: 1997. Pp. 433

£45, \$59.95

Ian Aitchison

A century can be a convenient unit of history. During the past hundred years or so, quantum field theory has replaced classical mechanics as the framework of fundamental physics. The climax of this development has been the spectacularly successful Standard Model, a relativistic quantum gauge field theory describing matter and forces at the smallest sizes currently accessible.

But the intricate theoretical structure has imperfections, and the general quantum field framework has failed to accommodate gravity on the same footing as the other known forces. There are tantalizing hints of a new physics beyond quantum field theory. It seems a good moment to take stock.

Tian Yu Cao's timely book provides a broad overview of twentieth-century field theories. The story is presented in three parts, each devoted to a general line of development which he subsumes under the heading of a "research programme".

The first part covers the geometrical programme (Einstein's metric field theory of gravitation, and its developments); the second

concerns the quantum field programme (the emergence, formulation and application of quantum field theory, including current algebra, anomalies, renormalization and the renormalization group); and the third describes the gauge field programme (the Standard Model, and some extensions). The discussion is on occasion necessarily technical, but this is not an advanced physics textbook: rather, it is an up-to-date, well informed and detailed historical account of the conceptual origins and development of twentieth-century field theories, structured in terms of the above programmes.

This is not all, however. Surrounding this substantial historical meat are introductory and closing chapters that are philosophical in character. For me, they are the most stimulating parts of the book.

Cao's aim is to show that there is a pattern in the conceptual evolution that he describes. In particular, he is concerned to combat the claim, originating in the work of Kuhn and others, that the history of science reveals no coherent direction of ontological development, and hence that no theoretical ontology can be taken as the real ontology of the world. Or, in other words, no matter how empirically successful our succeeding theories may be, they do not provide access to metaphysical reality. The contrasting position that Cao wishes to defend is characterized by him as "structural realism" which, roughly speaking, holds that the structural relations between entities (often expressed by mathematical structure) in a successful theory should be taken as real, rather than the entities themselves.

For this enterprise to succeed, Cao must show that structural properties of theoretical ontologies do persist across conceptual revolutions. This requirement, together with a desire to defend the rationality of scientific growth, leads him to propose that scientific revolutions can occur by what he calls "ontological synthesis", in which an outmoded ontology survives as an epiphenomenon, derivable from a later and more fundamental ontology. Like most physicists, I am all in favour of realism and rationality, but, while I applaud Cao's attempt, I think his analysis is open to serious challenge.

Classical field theory arose from the need to explain how physical actions, such as gravity and electromagnetism, can apparently be transmitted between bodies some distance from each other; this is in contrast to mechanical actions, which occur by direct contact. Newton hated the idea of action at a distance; and even Maxwell, whose equations for the dynamics of the electromagnetic field are now part of the Standard Model, thought his equations described the dynamics of a mediating medium, the ether, which he took to be some sort of mechanical substance. The ontological shift, by which the field became established as a non-mechanical entity in its



The magic of mushrooms

Life may be too short to stuff a mushroom, as the writer Shirley Conran said, but make time to read all about them in Elio Schaechter's funny and fascinating book, *In The Company of Mushrooms: A Biologist's Tale* (Harvard University Press, \$24.95, £16.50). Fungi have been found in fossilized wood 300 million years old, yet they are more closely related to animals

than to plants. And there's no shortage of them: Schaechter estimates that there are about two tons of fungi for every human being on Earth. Vital in the kitchen, fungi play an even more important role in recycling. "Life on earth as we know it would be impossible without them," he says. So we'll have to forgive them for Dutch elm disease and the Irish potato famine.

In retrospect chosen by John L. Casti

Gödel, Escher, Bach: An Eternal Golden Braid

by Douglas Hofstadter
(1979)

A classic dinner-party game I like to play is to ask my guests what science book published, say, during the past 25 years they would take if they were to be stranded indefinitely on a desert island. For me the choice is easy: I would grab a copy of Douglas Hofstadter's *Gödel, Escher, Bach* (*GEB*) before abandoning ship.

Alternating witty, enlightening dialogues with chapters on such meaty topics as self-reference, recursion, meaning in mathematics, formal systems of logic, brains, the weird and wonderful art of M. C. Escher, minds, thoughts, typesetting, Gödelian paradoxes, self-replication, artificial intelligence and hierarchies, *GEB* is more like a summary statement of most of modern cognitive science than a thematic book, more of a romp over the modern intellectual landscape than a scholarly monograph. Yet it manages to be all these things and more.

At the time *GEB* was sitting on bestseller lists around the world, a lot of people joked that it was a book everyone bought and no one read. Well, not me. I have a friend who says that each Christmas he rereads every Sherlock Holmes story because he learns something new with each rereading. I don't want to claim that I've now reread *GEB* 17 times, but I have read it cover-to-cover several times — and always with great profit. One reading taught me things about Bach fugues that I had never properly

appreciated, another time I gained a deeper insight into the figure-ground distinction, while a third reading heightened my sensitivities to the connection between Gödel's "incompleteness theorem", self-reference and Escher's famous engraving *Ascending and Descending*. All these intellectual gleanings are secondary matters to the book's main theme, which is an extended meditation on the possibility of duplicating human thought processes in a machine.

I don't think it would be too great a claim to date the birth of the 'connectionism' movement in artificial-intelligence research to the publication of *GEB* in early 1979. This is the view that the physical structure of the human brain matters when it comes to trying to duplicate in a computing machine what the brain does.

Before then, workers had spent the better part of 30 years trying to skim off human intelligence from the actual physiological structure of the brain, essentially ignoring things like behaviour and connective structures at the level of the brain's neurons. Hofstadter redirected attention to these lower-level structures, arguing that if one wants to capture human cognition in a mechanical intelligence, the machine must mirror these low-level features of the brain. *GEB*'s presence on the bestseller lists for months, along with its winning the 1980 Pulitzer prize for general nonfiction, brought this connectionist view of machine intelligence back to the forefront of research in artificial intelligence. Thus were reborn the theories of neural networks, genetic

algorithms, evolutionary programming and all the other 'hot' theories in the field.

It is interesting to ponder what makes *GEB* a book many purchasers have never read, while it has changed other folk's view of the world — like mine. On the downside, the book is a real doorstop, 778 pages long and more than two pounds in weight. So it's a bit of a package for comfortable reading on the subway or the beach. Some people have complained that it is rather disjointed and rambling, moving from topic to topic and back again with no apparent beginning or end. While I have to admit that there is some truth to these allegations, one might register the same complaints about the *Encyclopedia Britannica*. *GEB* cannot be measured by the standards usually employed to evaluate books. It's a one-of-a-kind volume that has to be looked at in its own right, which to their eternal credit the Pulitzer committee members immediately recognized.

On the asset side of the ledger, *GEB* is the best example I know of a 'transdisciplinary' book. Laying out explicit linkages between the humanities, art, science and mathematics, it proves that the landscape of the intellect does not come conveniently packaged into compartments such as English, physics, computer science and psychology. In some ways, *GEB* is an entire humanistic education between the covers of a single book. So, for my next visit to a desert island, give me sun, sand, water and *GEB*, and I'll live happily ever after.

John L. Casti is at the Santa Fe Institute, Santa Fe, New Mexico 87501, USA.

own right, is very drastic, and Cao does not explain how a mechanical ontology can emerge from field physics as an epiphenomenon. Perhaps we just have to accept an ontological enlargement.

The situation is much worse, I believe, in the case of quantum field theory. Superficially, quantum field theory sounds like a synthesis of quantum mechanics and classical field theory. But this immediately leads to a problem. The quanta of force fields are bosons, yet the theory must also include fermions, for which there is no classical field. As Cao explains, this led historically to the procedure called "second quantization", which seems to be associated with a particle rather than a field ontology. But either ontology (particle or field) is problematic, because bosonic and fermionic quantum fields are quantum operators whose matrix elements are quantities from which only probabilities can be calculated, as is typical in quantum mechanics.

Cao claims that physicists have still not clarified the "real material" versus "probabilistic" character of the quantum field. He himself regards it as a new kind of substantial field, and with this view it is easier to find the kind of ontological continuity he wants. For

example, he "can claim without hesitation that the gauge field programme is a direct descendant of the geometrical programme".

In fact, quantum field theory is really the result of trying to reconcile quantum mechanics with special relativity. Its ontology is no different from that of quantum mechanics itself — and here we hit a barrier. Ontology, Cao says, is "concerned with an autonomous existence without reference to anything external". But, according to what is still the standard (Bohrian) view, although the existence of some independent microscopic reality behind observed phenomena is accepted, questions about the nature or state of that reality have meaning only in the context of explicitly stated external experimental setups. In some way, attributes and quantities — even entities themselves — are ceaselessly fluctuating at the quantum level, and become real only as a result of 'measurement'. The simple connection between observed properties and ontological properties of an independently existing object has been severed. Of course, other interpretations have been proposed such as the 'many worlds' one; but this example has rather too much ontological baggage for many people's taste.

Perhaps Cao should have avoided the 'substantial' aspect of ontology, and stuck to the 'structural' aspects he himself advocates: witness, for example, the mathematical survival of Maxwell's equations. But, even at this level, the significance of the mathematical similarity between the space-time geometry of general relativity and 'internal' symmetries of gauge field theory is unclear.

Despite their having been invented in the same century, there is still a great divide between general relativity and quantum field theory. Perhaps the difficulty of reconciling them is an indication that our current ontological categories are inadequate, and that Cao's bold attempt at ontological synthesis is simply premature. □

Ian Aitchison is in the Department of Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK.

correction

The title of the book by Harrison G. Pope Jr reviewed by Stuart Sutherland in last week's issue (*Nature* **388**, 239; 1997) was given incorrectly. It is *Psychology Astray: Fallacies in Studies of 'Repressed Memory' and Childhood Trauma*.

Recombination of protein domains facilitated by co-translational folding in eukaryotes

William J. Netzer* & F. Ulrich Hartl*†

* Cellular Biochemistry & Biophysics Program and Howard Hughes Medical Institute, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

† Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, 82152 Martinsried, Germany

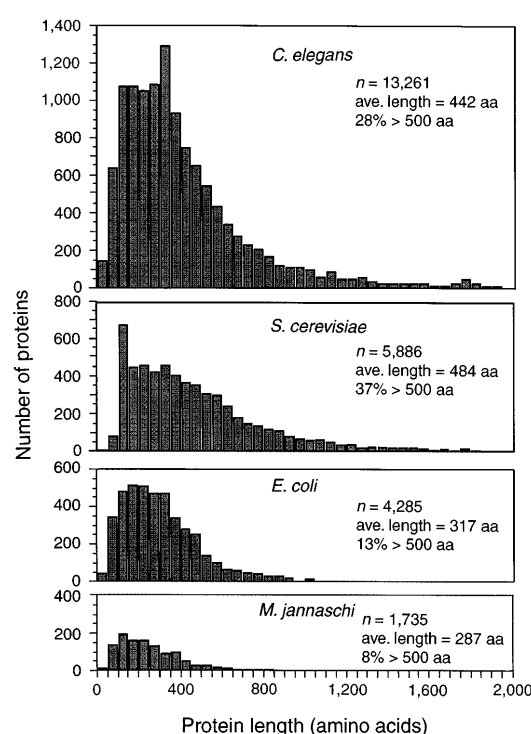
The evolution of complex genomes requires that new combinations of pre-existing protein domains successfully fold into modular polypeptides. During eukaryotic translation model two-domain polypeptides fold efficiently by sequential and co-translational folding of their domains. In contrast, folding of the same proteins in *Escherichia coli* is post-translational, and leads to intramolecular misfolding of concurrently folding domains. Sequential domain folding in eukaryotes may have been critical in the evolution of modular polypeptides, by increasing the probability that random gene-fusion events resulted in immediately foldable protein structures.

The evolution of eukaryotes has been characterized by a dramatic increase in the diversity and structural complexity of proteins. This has been attributed to the organization of eukaryotic genes into exons and introns, which is thought to have increased the probability of favourable gene duplication events as well as the generation of proteins with new functions by joining initially separate DNA sequences ('exon shuffling')¹⁻³. Proteins generated in this way typically contain two or more domains linked sequentially that are encoded either by distinct exons (in the case of more recently evolved structures⁴) or by blocks of exons². Importantly, domains (~100–300 amino acids in length) are not only units of function but also represent autonomous structural units of folding⁵⁻⁷. Independent of the precise genetic mechanism involved, combining ancestral protein domains that were originally separate is apparently the basic principle by which many polypeptides with new functions

evolved. This process has occurred with much higher frequency in eukaryotes than in prokaryotes, as is evident from a simple comparison of prokaryotic and eukaryotic protein structures using the information recently provided by several genome projects (Fig. 1). Although the size distribution of protein domains is uniform throughout the kingdoms of life², eukaryotes contain a much greater number (and proportion) of longer polypeptides combining multiple domains.

In vitro refolding studies indicate that the tendency of polypeptide chains to misfold increases significantly with their chain length, whereas spontaneous folding is usually efficient for small single-domain proteins⁸. Furthermore, expression of eukaryotic multi-domain proteins in bacteria frequently leads to aggregation⁹. One might therefore predict that eukaryotic cells have developed specific mechanisms to ensure the efficient folding of these proteins. For

Figure 1 A comparison of protein length distributions for *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *E. coli* and the archaeobacterium *Methanococcus jannaschi*. Protein number, based on genome analysis, is plotted against protein length (number of amino acids), where length windows correspond to 50 amino acids. The total number of proteins analysed per genome (*n*), the average polypeptide length and the fraction of proteins exceeding a length of 500 amino acids (aa) are given. The last 1% of proteins (>2,000 aa) are not included. Data are derived from the following sources and represent essentially complete genomes, except for *C. elegans*, for which about 95% of the genome is sampled here: *C. elegans*, Batch Entrez NCBI protein database, <http://www3.ncbi.nlm.nih.gov:80/cgi-bin/EntrezBatch/nph-batch/result>; *S. cerevisiae*, FTP://darcy.bu.edu/pub/genomes/; *E. coli* and *M. jannaschi*, TIGR Microbial Database (MDB) of the Institute for Genomic Research, <http://www.tigr.org:80/tdb/mdb/mdb.html>. The following individuals are acknowledged for their help in performing this analysis: M. Adams and O. White (Institute for Genomic Research, Rockville, MD), T. Smith and J. Freeman (The BioMolecular Engineering Center, Boston, MA), and L. Hillier (Washington University, St Louis, MO).



example, their folding might be mediated by the cellular machinery of molecular chaperones, in particular by the cylindrical chaperonins, which can prevent aggregation by sequestering a folding polypeptide within a central cavity¹⁰. However, the chaperonin of the eukaryotic cytosol seems to mediate the folding of only a specific subset of polypeptides^{11,12}. Moreover, the lumen of the endoplasmic reticulum, the site of folding of many composite polypeptides, lacks a chaperonin altogether.

Here we tested the hypothesis that sequential folding of domains during synthesis on ribosomes minimizes the error rate of folding for complex proteins. This mechanism would reduce the problem of folding a large polypeptide to that of folding small polypeptide modules. We have constructed new composite proteins by joining single-domain proteins with a flexible linker. *In vitro* refolding of these fusion proteins is inefficient as a result of intramolecular misfolding. In contrast, during eukaryotic translation, the model proteins reach their native state efficiently by sequential and co-translational folding of domains. Remarkably, folding of the same proteins in a bacterial system is post-translational and leads to aggregation. We propose that co-translational, domain-wise folding is the basis for the efficient folding of a large number of eukaryotic multidomain proteins.

Refolding of two-domain proteins *in vitro*

A fusion protein was constructed consisting of two monomeric, single-domain cytosolic proteins, human H-Ras (M_r 21K) and mouse dihydrofolate reductase (DHFR, M_r 20K). The domain folds of these proteins exist also in prokaryotes. The carboxy terminus of Ras was joined to the amino terminus of DHFR by means of a flexible polypeptide linker containing Ser-Gly-Ile

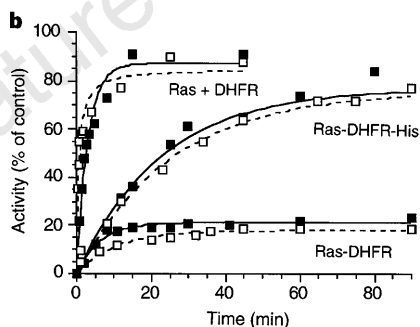
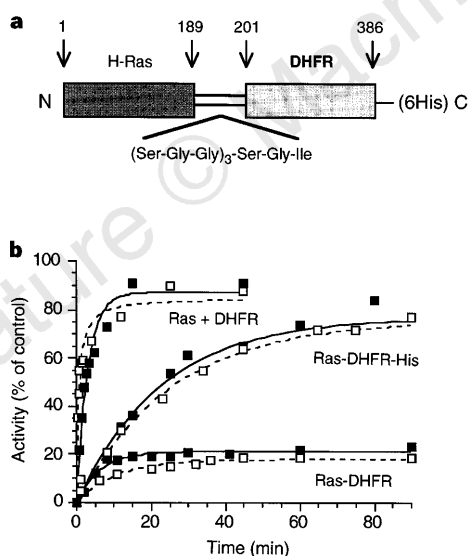


Figure 2 Refolding *in vitro* of Ras-DHFR fusion proteins. **a**, Diagram of the synthetic fusion protein, Ras-DHFR. The C terminus of human H-Ras is linked covalently to the N terminus of mouse DHFR through a 12-residue flexible peptide linker¹³. A second version of the fusion protein (Ras-DHFR-His) contains a six-histidine tag at the C terminus. **b**, Kinetics and yields of refolding upon dilution from 6 M GdmCl into refolding buffer at 25 °C for Ras-DHFR, Ras-DHFR-His, and mixtures of Ras and DHFR. Final protein concentrations were 0.5 μ M for the fusion proteins and for Ras and DHFR. Activities (³H-GDP binding for Ras (filled squares) and enzymatic activity for DHFR (open squares)) are expressed as percentage of native, single-domain Ras and DHFR controls. Refolding kinetics of DHFR and DHFR-His do not differ significantly, so only the kinetics of refolded DHFR is shown. Refolding of mixtures of Ras and DHFR were indistinguishable from folding of each protein in isolation. Data were fitted to first-order exponentials.

repeats¹³ (Fig. 2a). Similar linkers join the domains in many eukaryotic proteins, including those that function in signal transduction. Two fusion proteins (with and without a C-terminal six-histidine (His) tag) were expressed in bacteria, mainly as inclusion bodies. The purified proteins were denatured in 6 M guanidinium-HCl, and refolding was conducted by rapid dilution into a refolding buffer. Refolding of Ras-DHFR-His was about 80% efficient for both domains (Fig. 2b). However, Ras-DHFR (minus the His tag) folded only with 20% efficiency. Native fusion protein was efficiently cleaved in the linker region by proteinase K (see below). The protease resistance of the Ras and DHFR domains within the folded fusion protein was indistinguishable from that of the unlinked proteins. In contrast, attempted refolding of Ras-DHFR produced mostly misfolded protein that was readily digested in its non-aggregated state (data not shown).

Refolding of Ras and DHFR was analysed at submicromolar protein concentrations where sedimentable aggregates did not form. Mixing the two single-domain proteins in their unfolded states (in *trans*) had no effect on the yield and rate of refolding (Fig. 2b). The time constants (τ) for folding of the ras and DHFR proteins were 3 and 1.7 min, respectively. (Refolding rates for DHFR and DHFR-His were indistinguishable). However, in the His-tagged fusion protein both Ras and DHFR domains (in *cis*) refolded concurrently with an 8-fold slower time constant of $\sim$ 24 min, suggesting that there is intramolecular interference between the Ras and DHFR polypeptides.

In contrast to Ras-DHFR-His, refolding of Ras-DHFR occurred with low yield, perhaps because of intermolecular aggregation or intramolecular misfolding. Aggregation is a multimolecular process, but intramolecular misfolding is unimolecular and is therefore

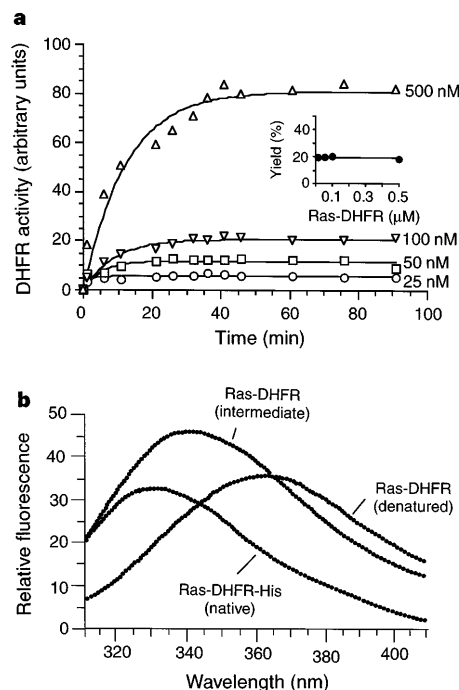


Figure 3 Partitioning of *in vitro* refolded Ras-DHFR into native and non-native folds. **a**, Refolding of Ras-DHFR at varying final concentrations measured as the increase in DHFR activity over time (see Fig. 2b). Insert shows concentration independence of refolding yields. **b**, Tryptophan fluorescence spectra excited at 295 nm for refolded Ras-DHFR (intermediate) in comparison to native Ras-DHFR-His and denatured Ras-DHFR in 6 M GdmCl (0.5 μ M final protein concentration). The six-histidine tag did not influence the fluorescence signal.

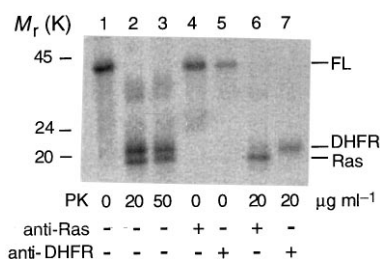


Figure 4 Efficient folding of Ras-DHFR upon translation in reticulocyte lysate. Formation of ^{35}S -methionine-labelled protease-resistant Ras and DHFR domains on translation in reticulocyte lysate detected by immunoprecipitation. Translation was for 40 min at 25°C in a non-synchronized reaction. Lanes 1–3, translation product was incubated in the absence or presence of proteinase K (PK, 20 and 50 $\mu\text{g ml}^{-1}$); lanes 4–7, immunoprecipitations with anti-Ras and anti-DHFR antibodies. Analysis by 12.5% SDS-PAGE and phosphorimaging. FL, full-length protein.

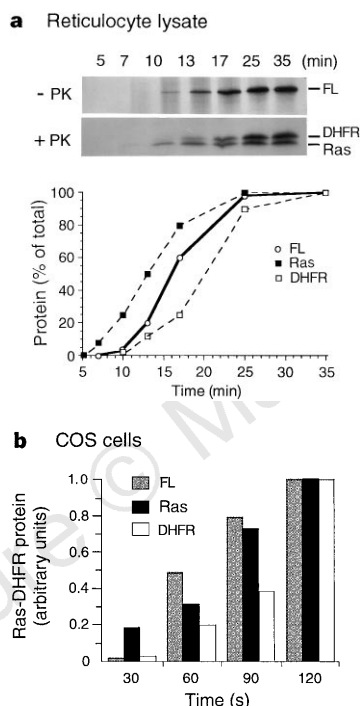


Figure 5 Co-translational folding of Ras-DHFR in the eukaryotic cytosol. **a**, Synchronized translation of Ras-DHFR in rabbit reticulocyte lysate. Reactions were stopped by dilution into ice-cold buffer containing cycloheximide either with or without proteinase K (+PK and -PK) and analysed by SDS-PAGE and fluorography (top). FL, full-length protein. Note that the protease-resistant Ras domain contains only half as many methionines as the DHFR domain. Data from synchronized translations were quantified by densitometry (average of three independent experiments) and corrected for the differences in methionine contents in Ras and DHFR. **b**, Expression of Ras-DHFR in COS cells. Transfected COS cells were labelled at 37°C with ^{35}S -methionine. Cells were lysed in ice-cold buffer containing digitonin or digitonin and PK. Polypeptides were immunoprecipitated with anti-Ras and anti-DHFR antibodies and analysed by SDS-PAGE and phosphorimager quantification. (Expression levels of endogenous Ras and DHFR are less than 10% of that of the fusion protein.) Emergence of each protein species is shown as the fraction of the amount detected at 120 s of labelling when the relative amounts of the different proteins were constant. The approximate rate of polypeptide synthesis was calculated by measuring the time required after adding ^{35}S -methionine to cells to achieve a constant proportion of labelled, proteolysed Ras domain relative to labelled, full-length product. The time required to aminoacylate tRNA was taken to be 15–30 s.

concentration independent. The fraction of protein that reached the native state remained the same ($\sim 20\%$) at increasing concentrations of fusion protein in the range 25–500 nM, indicating that intermolecular aggregation was not governing the efficiency of refolding (Fig. 3a). Dynamic light scattering determined that the non-native intermediate, once formed, remained monomeric (M_r 43K) up to low micromolar concentrations (data not shown). The intermediate had the following physical properties: a maximum tryptophan fluorescence emission at 340 nm (compared with 331 nm for native Ras-DHFR-His), consistent with a collapsed conformation in which Trp residues are incompletely buried (Fig. 3b); the presence of significant α -helical structure by far-ultraviolet circular dichroism spectroscopy; enhanced binding of the fluorescent probe 8-anilidonaphthalene-1 sulphonate compared to the native protein; and hardly any cooperativity of unfolding (data not shown).

In summary, during refolding of Ras-DHFR, the two domains interfere intramolecularly, resulting in the accumulation of a non-native intermediate(s). The presence of the histidine tag seems to destabilize this intermediate such that slow folding to the native state occurs.

Co-translational folding in eukaryotic cytosol

Remarkably, the Ras-DHFR fusion protein folded with almost full efficiency on translation in a rabbit reticulocyte lysate. Folding was assayed by partial proteinase K digestion followed by immunoprecipitation of Ras and DHFR domains (Fig. 4). Nearly all newly synthesized protein was soluble and was cleaved in the linker into Ras and DHFR fragments, the signature of the native fusion protein.

Next we determined the timing of domain folding during synthesis by synchronizing translation with aurintricarboxylic acid, which inhibits reinitiation of translation¹⁴ and allows only a single round of elongation (Fig. 5a). The time required to synthesize a single molecule of Ras-DHFR was approximately 9 min at 24°C. Formation of protease-resistant Ras domain occurred consistently before the full-length fusion protein, indicating that Ras folded co-translationally. The folded DHFR domain appeared only after synthesis of full-length protein was completed, consistent with the notion that folding of a single protein domain generally requires its complete extrusion from the ribosome¹⁵. Equimolar amounts of Ras and DHFR bands were present at late time points of translation, demonstrating that the fragments resulted from digestion of productively elongating polypeptide chains (note that DHFR contains eight methionines, compared with just four in the protease-stable Ras domain). Thus, in contrast to refolding *in vitro*, during cell-free translation folding of Ras-DHFR fusion protein occurs in a domain-wise manner and is highly efficient. Attempts to detect an interaction of translating Ras-DHFR with the eukaryotic chaperonin TRiC were unsuccessful (data not shown), suggesting that co-translational folding of Ras-DHFR is independent of chaperonin.

We then investigated whether co-translational folding of Ras-DHFR occurred *in vivo* and at 37°C. Folding was assayed in COS cells transiently transfected with the fusion construct (Fig. 5b). It was not feasible to produce a synchronized translation *in vivo*. However, we reasoned that if co-translational folding were to occur, it could be observed after short times of labelling. The approximate rate of polypeptide synthesis was calculated as one Ras-DHFR polypeptide per 120 s, corresponding to an elongation rate of ~ 3 amino acids per second (see Fig. 5b). Radiolabelling of cells was stopped after various times by dilution into ice-cold buffer containing cycloheximide and digitonin with or without proteinase K. Folded Ras domain was seen in immunoprecipitations after just 30 s of labelling, at a time when very little labelled full-length product and folded DHFR domain were detectable (Fig. 5b). Based on the rate of translation, any labelled, full-length Ras-DHFR occurring at that time could be labelled only in the DHFR domain. Therefore, any labelled, protease-resistant Ras appearing at this time must be

derived from protein that folded before elongation was completed. This folded Ras is likely to represent productively translating protein, because the proportion of Ras domain to full-length protein at later time points corresponded to the composition of the fusion protein (data not shown). Thus co-translational, domain-wise folding occurs for Ras-DHFR under physiological conditions *in vivo*. Furthermore, virtually all of the protein expressed in COS cells was folded and soluble (Table 1).

Post-translational folding in bacterial cytosol

Only 10% of total Ras-DHFR-His was soluble and native when expressed in *E. coli* (at 25–37°C), independent of the time of induction. In contrast, no folded Ras-DHFR (minus His tag) was detected under the same conditions of expression (Table 1). Because the C-terminal His tag could only influence a post-translational folding mechanism, it was likely that folding in *E. coli* might occur after the fusion polypeptide is released from the ribosome.

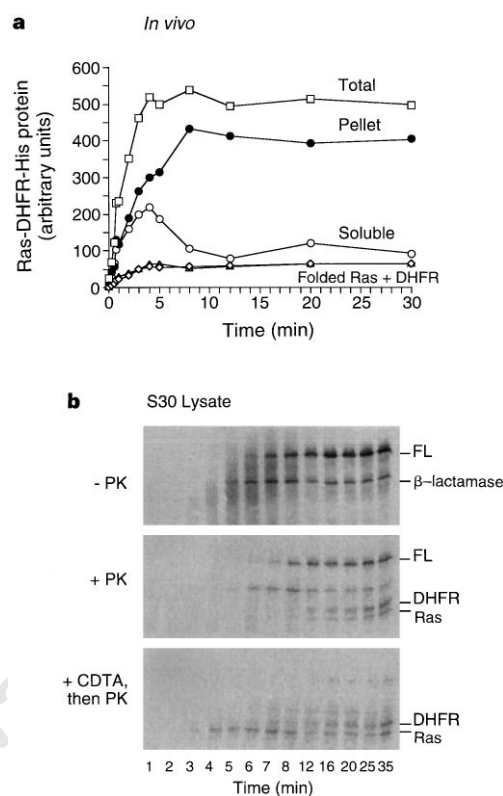


Figure 6 Post-translational folding of Ras-DHFR-His in *E. coli* cytosol. **a**, Expression of Ras-DHFR-His in *E. coli* *in vivo*. Time course of production of full-length protein in total (open squares), soluble fraction (open circles) and insoluble pellet (filled circles), and of soluble, proteinase K-resistant Ras (open diamonds) and DHFR (open triangles) domains. Spheroplasts were pulse-chase labelled at 30°C with ³⁵S-methionine. Amounts of Ras are corrected for methionine content. Analysis by SDS-PAGE and phosphorimager quantification. **b**, Synchronized translation of Ras-DHFR-His and β-lactamase (from the same plasmid) at 25°C in *E. coli* S30 lysate. Reactions were stopped by dilution into ice-cold buffer containing chloramphenicol (top); chloramphenicol and proteinase K (PK) (middle); or chloramphenicol, CDTA and PK (bottom). In the bottom panel, PK was added 5 min after CDTA. FL, full-length Ras-DHFR-His. Analysis was by SDS-PAGE and phosphorimaging. CDTA treatment was correlated with the release of nascent chains from ribosomes, as demonstrated by sedimentation analysis. Note that β-lactamase is synthesized as a precursor containing a hydrophobic signal sequence. Instability of β-lactamase in the presence of PK may be enhanced by binding to chaperones, such as GroEL.

To explore this possibility, viable spheroplasts prepared from *E. coli* cells expressing Ras-DHFR-His were labelled with ³⁵S-methionine at 30°C, followed by unlabelled methionine after 15 s (Fig. 6a). Ras-DHFR-His was the predominant polypeptide species synthesized. Further incorporation of radiolabel into newly synthesized polypeptide was effectively prevented after 2 min of chase (data not shown), but production of labelled full-length protein continued owing to completion of elongating chains (Fig. 6a). At different times spheroplasts were rapidly cooled and lysed with digitonin in the presence or absence of proteinase K. As soon as full-length protein was synthesized about 50% of the protein aggregated and was recovered in the pellet fraction. This material was protease resistant (data not shown). The remainder was soluble and protease sensitive, but only ~25% was stably folded, as indicated by cleavage in the linker region into protease-resistant Ras and DHFR domains. Furthermore, the two domains folded concurrently, and only after full length protein had been synthe-

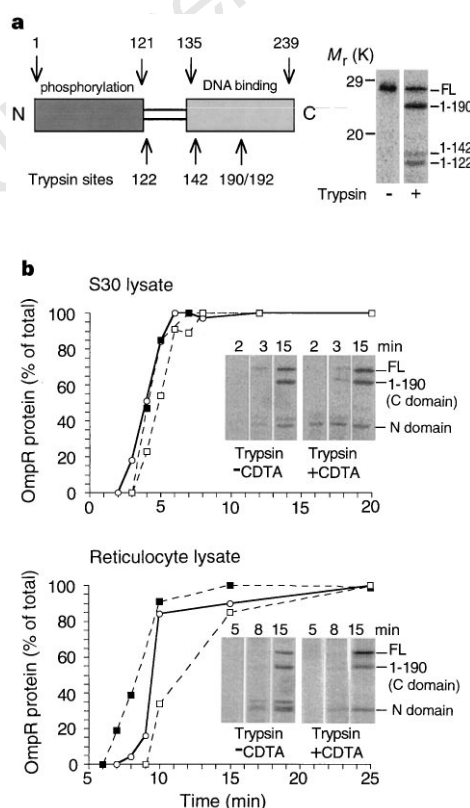


Figure 7 Post-translational and co-translational folding of *E. coli* OmpR. **a**, Left, schematic representation of the two-domain structure of OmpR. The linker segment has the sequence RQANELPGAPSQEE (single-letter code starting with residue 122)^{18,19}. Arrows indicate the positions of the major trypsin cleavage sites in the folded protein. Right, SDS-PAGE of OmpR after 20 min translation in S30 followed by trypsin treatment. FL, full-length OmpR; fragment 1-190 results from trypsin cleavage in the folded C-terminal domain. It consists of two peptides encompassing residues 1-190 and 1-193 (ref. 18). The two lower fragments correspond to the N-terminal domain. **b**, Synchronized translation of OmpR in *E. coli* S30 lysate (top) and in reticulocyte lysate (bottom). Analysis was as in Fig. 6b and Fig. 5b, respectively, except that trypsin was used to monitor folding. Amounts of full-length OmpR (open circles), as well as trypsin-resistant fragments 1-121 plus 1-142 (N domain) (filled squares) and of fragment 1-190 (C domain) (open squares), were quantified and are expressed as a percentage of the amounts present at 20 (S30) or 25 min (reticulocyte lysate) of translation. At the times indicated, translation reactions were incubated with or without CDTA (5 min at 25°C in the presence of chloramphenicol) followed by trypsin. Insets show analysis by SDS-PAGE. Note that the speed of polypeptide elongation is very similar in the two systems, although initiation of translation is faster in bacterial S30.

sized (Fig. 6a). During the chase period a second, slower phase of protein aggregation occurred from the soluble pool of Ras–DHFR molecules that did not reach the native state (Fig. 6a). These results suggested that folding of the fusion protein occurred by a post-translational mechanism and was inefficient as a result of intramolecular misfolding.

Translation in *E. coli* S30 lysate was used for further investigation. Compared with expression in intact *E. coli*, Ras–DHFR–His synthesized in S30 folded with a higher efficiency of 30% (Fig. 6b and Table 1). The remainder of the full-length protein was relatively protease resistant, suggesting the formation of misfolded or aggregated species as *in vivo*. (Almost no folded Ras or DHFR domain was produced from Ras–DHFR (Table 1).) A synchronized translation confirmed that folding of Ras–DHFR–His was indeed post-translational (Fig. 6b, top and middle panels), in marked contrast to the observations in reticulocyte lysate. The speed of translation in S30 and reticulocyte lysate, and also the physical conditions of translation (pH, protein and salt concentrations), are very similar. Furthermore, the amount of protein synthesized in S30 is small compared to the amount *in vivo*, and inclusion bodies are absent. Thus, misfolding on translation does not seem to be dependent on the physiological speed of *E. coli* translation, nor on the concentration of the newly synthesized polypeptide or the presence of pre-existing aggregates.

To test whether it is the association with translating *E. coli* ribosomes that prevented polypeptide chains containing a completed Ras domain from folding, the Mg^{2+} -chelator CDTA was added to release ribosome-bound chains¹⁶ before digestion with proteinase K (Fig. 6b, bottom). This demonstrated that, before the appearance of full-length protein, nearly all of the Ras domain synthesized (but still ribosome bound) was foldable. Thus, during ongoing translation in the absence of CDTA, folding of Ras is delayed until full-length protein is synthesized. The proportion of foldable Ras was maximal when ribosome release was induced before synthesis of the C-terminal DHFR. At later time points the amount of foldable Ras decreased concomitantly with the increase observed in protease-resistant, apparently misfolded full-length protein (Fig. 6b, middle). This effect is very similar to the misfolding of fusion protein during refolding *in vitro*. (CDTA treatment reduced the protease resistance of misfolded full-length protein, presumably by destabilizing protein aggregates, but had no effect on the amount of folded Ras and DHFR.)

Chelating Mg^{2+} is expected to stabilize the interaction of unfolded polypeptide chains with ATP-dependent chaperones such as DnaK (Hsp70) and GroEL^{15,17}, rendering bound polypeptides sensitive to digestion by protease. It is therefore unlikely that these components are responsible for the prevention of co-translational folding of the Ras domain. Moreover, increasing the concentrations of chaperones in S30 lysate by 5- to 10-fold by adding purified DnaK, DnaJ, GrpE, GroEL and GroES did not affect the efficiency or timing of folding (data not shown). Thus the failure of the bacterial system in folding the fusion proteins co-translationally is related either to properties of the bacterial ribosome itself, or to components of the translation/folding machinery that interact tightly with nascent polypeptide chains, but dissociate upon chain release from the ribosome.

Post- and co-translational folding of OmpR

To determine whether post-translational folding in *E. coli* lysate is a unique property of Ras–DHFR, we performed similar experiments using the natural bacterial transcription factor OmpR (M_r 27K). OmpR consists of an N-terminal phosphorylation domain and a C-terminal DNA-binding domain joined by a ~12-residue flexible linker^{18,19} (Fig. 7a). Analysis of the folded structure of OmpR has been established using limited trypsin digestion¹⁸. Only a few of the 34 potential trypsin cleavage sites are accessible in the native protein. Cleavage at a major site in the C-terminal domain yields a fragment of residues 1–190 containing the entire N-terminal

Table 1 Folding efficiencies upon synthesis (% of total protein)

	Ras	DHFR	Ras–DHFR	Ras–DHFR–His
<i>E. coli</i>	>90*	>90	<2	10
S30 lysate	>90	>90	<5	30
COS cells	n.d.	n.d.	>90	n.d.
Reticulocyte lysate	>90	>90	>90	>90

Folding efficiencies of Ras, DHFR and Ras–DHFR fusion proteins upon synthesis *in vivo* and *in vitro* are expressed as the fraction of soluble protein having native structure as confirmed by limited proteolysis and DHFR activity measurements (for *E. coli* only). Data are averages of 3–6 independent experiments.

* Ras expression in *E. coli* was also determined in ref. 39.

domain, the linker and part of the C-terminal domain (Fig. 7a). Its appearance signifies that both the C-terminal and the N-terminal domains are folded. The two lower fragments (residues 1–122 and 1–142) are derived from cuts in or close to the linker region (occurring with lower efficiency) and contain the N-terminal domain plus part of the linker¹⁸. Appearance of these fragments alone indicates that only the N-terminal domain has folded. The remaining C-terminal segment (residues ~142–190) is not resolved on the gel.

Folding of OmpR was analysed during translation in S30 and reticulocyte lysates (Fig. 7b). Efficient folding in the bacterial lysate was similar to that of Ras–DHFR–His in that folded domains were produced only after synthesis of full-length protein (Fig. 7b, top). Again, CDTA-induced release from ribosomes showed that folding of the N-terminal domain was retarded as long as the protein was associated with ribosome (Fig. 7b, inset). Nevertheless, the N-terminal domain folded slightly faster than the C-terminal domain. An intrinsically slower rate of folding of the C-terminal domain could be an adaptation of OmpR to allow productive post-translational folding. Alternatively, the N-terminal domain may initiate folding co-translationally, but reaches its native, protease-resistant state only after synthesis of full-length polypeptide. Because the speed of bacterial translation *in vivo* is much more rapid than in S30 lysate, this would strongly favour a completely post-translational folding mechanism of OmpR. Remarkably, domain folding of OmpR was sequential and co-translational in reticulocyte lysate (Fig. 7b, bottom). Addition of CDTA before trypsin treatment did not increase the amount of folded N domain (Fig. 7b, insert), indicating that its folding was not retarded by the eukaryotic ribosome. These results confirm that the eukaryotic and bacterial translation systems differ in their ability to support the co-translational folding of modular proteins.

Discussion

Our observations provide evidence for the biological significance of co-translational polypeptide chain folding. Rapid and efficient folding of modular model proteins in the eukaryotic system depends on sequential folding of their domains during synthesis. Strikingly, the bacterial translation system lacks the capacity for folding these model proteins co-translationally, and produces misfolded species that form inclusion bodies.

There are estimated to be only about 1,000 protein modules with unique three-dimensional structures^{7,20}. Based on our observations, we propose that a shift from post-translational to co-translational folding mechanisms has occurred during evolution. This may have provided the basis for the efficient exploitation of a limited set of available structures to evolve modular polypeptides with new functions through gene fusion events (the ‘folding shift hypothesis’). Alternatively, the restricted capability of prokaryotes to fold these proteins, along with their streamlined genomic organization, may explain the limited complexity of prokaryotic proteins. In the folding shift hypothesis we do not specify whether the ability to support co-translational domain folding has been lost (or reduced) by prokaryotes (perhaps together with split genes²), or whether it has been acquired by eukaryotes. Our distinction

between prokaryotic and eukaryotic folding does not predict that bacteria are unable to generate immediately foldable (and thus selectable) proteins from random gene fusions, but that such events would be rarer in organisms that do not generally use co-translational folding.

Co- versus post-translational folding. Co-translational folding has been observed in eukaryotic systems^{15,21–23}, but has not previously been identified as a mechanistic requirement for successful folding of modular proteins. Independent support for the predominance of post-translational folding in bacteria comes from our finding that most of the polypeptides synthesized in the *E. coli* cytosol can associate with the chaperonin GroEL before folding (unpublished observations), although GroEL does not recognize ribosome-bound chains^{10,24}. The physical basis for the difference between the prokaryotic and eukaryotic translation and folding machineries remains to be defined. Generally, the higher speed of bacterial translation (~15 residues per second²⁵) would favour post-translational folding. However, post-translational folding was observed even when the speed of translation was reduced *in vitro* to that of the eukaryotic system. Prevention of folding of ribosome-bound protein in *E. coli* may be caused by either molecular chaperones or by the bacterial ribosome itself.

Although we do not exclude the possibility that co-translational folding may occur for certain bacterial proteins (for example, exceptionally large polypeptides such as β -galactosidase), it is plausible that the selection (or maintenance) of co-translational folding mechanisms may not have been stringent in bacteria, where rapid polypeptide elongation is geared towards short generational times. Present-day bacterial proteins may be selected to fold rapidly when released from the ribosome. This may explain why the solubility of small eukaryotic proteins in *E. coli* sometimes increases upon their translation as fusions with a bacterial 'carrier' polypeptide (or with glutathione S-transferase), independent of whether the carrier is fused to the N terminus or the C terminus of the eukaryotic protein^{26,27}. On the other hand, the inefficient folding of many modular eukaryotic proteins in *E. coli*⁹ may result from the basic inability of bacteria to fold these proteins co-translationally. Another effect of co-translational folding in eukaryotes may be the expression of enzymes cooperating in specific metabolic pathways as large, multidomain proteins, whereas in prokaryotes these activities are synthesized as separate polypeptides²⁸.

The role of molecular chaperones. The bacterial chaperonin acts in the post-translational folding of a wide variety of cytosolic proteins^{10,29}, but a general chaperonin seems to be missing from the eukaryotic cytosol. Unlike GroEL, the eukaryotic chaperonin TRiC is thought to be involved in the folding of only a restricted subset of cytosolic proteins, including actin and tubulin^{11,12}. It is conceivable that co-translational folding might have eliminated the selective pressure for a more generalized chaperonin activity by reducing the formation of kinetically trapped, aggregation-sensitive intermediates. Actin is composed of domains that consist of discontinuous amino-acid sequences³⁰, and, although multidomain, would not be expected to fold co-translationally. Proteins with similar architecture may have a strong dependence on TRiC for folding.

Many translating chains in eukaryotic cells seem to interact with Hsp70, however^{31,32}. The Hsp70 chaperone system is thought to prevent the aggregation of nascent polypeptide chains synthesized on polyribosomes. This may be critical before complete synthesis of a domain (and its extrusion from the ribosome) that is required for stable folding¹⁰. ATP-dependent binding and release cycles of Hsp70 and co-chaperones may be sufficient to mediate the co-translational folding of modular proteins, whereas shielding within the chaperonin cavity during folding may be needed predominantly by proteins with structurally unstable (or discontinuous) domains that are unable to fold independently. Thus, the capacity of eukaryotic cells to fold complex, modular proteins co-translation-

ally may have simplified the chaperone requirement of *de novo* protein folding. □

Methods

Plasmid construction. Human H-Ras (Val 12) cDNA and mouse DHFR cDNA were cloned in *E. coli* TG1 from pSP64 and pGEM-4 (Promega), respectively. Ras cDNA and DHFR cDNA were combined in a three-way ligation with a synthetic DNA linker encoding Val-Leu-Ser (the C-terminal tripeptide of Ras)-Gly-Gly-(Ser-Gly-Gly)₂-Ser-Gly-Ile-Met¹³. One clone was sequenced and identified as pSP64/pGEM-Ras-DHFR. Ras-DHFR and Ras-DHFR-His were cloned into pET3A (Novagen) using the polymerase chain reaction (PCR) and pSP64/pGEM4-Ras-DHFR as template. Ras-DHFR was also cloned into plasmid pcDNA3 (Invitrogen) for transient transfection of COS cells. Additionally, H-Ras (Val 12) and mouse DHFR were cloned into pET 3A for expression in *E. coli*.

Protein purification. Pet3A-Ras-DHFR-His was expressed in *E. coli* strain BL21 by IPTG induction (3 h at 37°C). Protein was purified to >95% on Ni-NTA³³ either from the soluble fraction or after dissolving insoluble protein in 6 M guanidinium-HCl (GdmCl). Purified protein was transferred into buffer A (20 mM Tris HCl, pH 7.4, 80 mM KOAc, 1 mM MgOAc, 2 mM DTT) containing 6 M GdmCl in case of protein purified from inclusion bodies. Protein concentrations were based on measurements of absorption at 280 nm (A_{280}). pET-3A-Ras-DHFR was induced as above. Inclusion bodies were washed, dissolved in 6 M GdmCl and purified to >80% by reverse-phase HPLC (C4, VYDAC column) over a 30–60% gradient of acetonitrile/0.1% trifluoroacetic acid. When native Ras-DHFR-His (previously purified on Ni-NTA) was subjected to the same purification scheme, no differences in specific activity of refolded Ras-DHFR-His were observed compared to Ras-DHFR-His purified by Ni-NTA alone. DHFR-His and DHFR (mouse) were expressed in soluble form and purified on Ni-NTA or methotrexate-agarose³³, respectively.

In vitro refolding. Protein dissolved in 6 M GdmCl was rapidly diluted 100-fold into refolding buffer (20 mM Tris-acetate, pH 7.4, 80 mM KOAc, 1 mM MgOAc, 2 mM DTT) containing protein ligands (50 μ M dihydrofolic acid or 2–5 μ M GDP). Refolding of DHFR was measured in refolding buffer containing 50 μ M dihydrofolic acid and 50 μ M NADPH by following the decrease in A_{340} over time¹⁷. Ras renaturation was measured in the presence of 2 μ M ³H-GDP (1.0 mCi ml⁻¹, NEN), and bound ³H-GDP was determined³⁴.

In vitro transcription and translation. *In vitro* transcriptions for translation in rabbit reticulocyte lysate were performed using pSP64/pGEM-4-Ras-DHFR linearized with BglII as template and SP6 polymerase (Promega). Translations were carried out in 50–70% nuclease-treated reticulocyte lysate (Promega) in the presence of 0.8 mCi ml⁻¹ ³⁵S-methionine (>1,000 Ci mmol⁻¹, Amersham) at 25°C (ref. 15). Coupled *in vitro* transcription/translation in *E. coli* S30 lysate³⁵ was carried out at 25°C with pET3A Ras-DHFR, pET3A Ras-DHFR-His, and the *E. coli* OmpR gene (inserted in pJES307, a derivative of pT7-7; a gift from A. West)³⁶ as templates in the presence of 0.6 mCi ml⁻¹ ³⁵S-methionine, 0.06–0.1 mg ml⁻¹ plasmid DNA, 2 μ g ml⁻¹ rifampicin, and 1.5 units ml⁻¹ T7 RNA polymerase. Translations were synchronized by adding 100 μ M aurintricarboxylic acid (Sigma) 3–9 min after addition of RNA (reticulocyte lysate) or 2.5–3 min after addition of DNA (S30 lysate)¹⁴.

For limited proteolysis, 3–5 μ l of translation reaction was diluted 50-fold in ice-cold buffer A (minus DTT) containing 5 mM MgOAc, 2 mM unlabelled methionine and 2 mM cycloheximide (reticulocyte lysate) or 10 μ g ml⁻¹ chloramphenicol (*E. coli* S30). Proteinase K (Boehringer) was present at 10 or 20 μ g ml⁻¹ and trypsin (Boehringer) at 10 μ g ml⁻¹, and incubation proceeded on ice for 10 min. Proteinase K was inhibited by addition of 1 mM phenylmethylsulphonylfluoride (PMSF) and trypsin by 2 molar equivalents of soybean trypsin inhibitor. When used, CDTA (10 mM) was also present during proteolysis. In several experiments the original Mg²⁺ concentration of the buffer was restored after CDTA induced dissociation of ribosomes (before addition of protease) to demonstrate that protease activity was not Mg²⁺ dependent. Proteolysed samples were precipitated by trichloroacetic acid (TCA) and analysed by SDS-PAGE.

Expression and labelling of Ras-DHFR *in vivo*. Transfected COS cells³⁷ were suspended at ~10⁶ cells ml⁻¹ in Dulbecco's modified Eagle medium lacking methionine and incubated for 30 min at 37°C. Cells (800 μ l) were pulse

labelled with 1.7 mCi ml⁻¹ ³⁵S-methionine. Portions of 50 µl were removed at times after addition of label and mixed in 200 µl ice-cold buffer A (minus DTT) containing 2 mM cycloheximide and 0.5% digitonin with or without 5 µg ml⁻¹ proteinase K (see above). Reactions were centrifuged at 15,800g at 4 °C for 10 min (pellets did not contain Ras–DHFR). TCA precipitates of supernatants were dissolved in 2% SDS buffer. Portions were diluted 1:10 in 10 mM Tris acetate, pH 7.4, 0.3 M NaCl, 5 mM EDTA, and 1% Triton X-100, containing either 10 µg ml⁻¹ anti-Ras monoclonal antibody (pan Ras AB-3; Oncogene Science), or 1/10 volume of anti-DHFR antiserum from rabbit, and incubated for 1 h on ice, followed by addition of protein A-Sepharose.

PET3A–Ras–DHFR–His was expressed in *E. coli* BL21 by IPTG induction (2 h at 30 °C). Before spheroplasting³⁸, cells were grown at 30 °C in M63 minimal medium lacking methionine (doubling time ~90 min). Spheroplasts were incubated with 60 µCi ml⁻¹ ³⁵S-methionine at 30 °C followed by addition of 1 mM unlabelled methionine after 15 s. Incorporation of label (TCA-precipitable radioactivity) stopped completely 2 min after adding unlabelled methionine. Portions were diluted 2-fold into ice-cold buffer A (minus DTT)/0.5% digitonin (lysis buffer) with or without proteinase K. After inhibition of proteinase K, reactions were separated into pellets and supernatants (5 min, 15,800g at 4 °C). Pellets were washed in lysis buffer and analysed either by SDS–PAGE or by immunoprecipitation for Ras and DHFR proteins.

Received 24 March; accepted 9 June 1997.

- Gilbert, W. The exon theory of genes. *Cold Spring Harb. Symp. Quant. Biol.* **111**, 901–905 (1987).
- Blake, C. C. F. Exons and the evolution of proteins. *Int. Rev. Cytol.* **93**, 149–185 (1985).
- Doolittle, R. F. & Bork, P. Evolutionary mobile modules in proteins. *Sci. Am.* **269**, 50–56 (1993).
- Stolzfus, A., Spencer, D. F., Zuker, M., Logsdon, J. M. & Doolittle, W. F. Testing the exon theory of genes: The evidence from protein structure. *Science* **265**, 202–207 (1994).
- Richardson, J. S. Describing patterns of protein tertiary structure. *Methods Enzymol.* **115**, 341–380 (1985).
- Doolittle, R. F. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64**, 287–314 (1995).
- Orengo, C. A., Jones, D. T. & Thornton, J. M. Protein superfamilies and domain superfolds. *Nature* **372**, 631–634 (1994).
- Jaenicke, R. Folding and association of proteins. *Prog. Biophys. Mol. Biol.* **49**, 117–237 (1987).
- Marston, F. A. O. The purification of eukaryotic polypeptides synthesized in *Escherichia coli*. *Biochem. J.* **240**, 1–12 (1986).
- Hartl, F. U. Molecular chaperones in cellular protein folding. *Nature* **381**, 571–580 (1996).
- Kubota, H., Hynes, G. & Willison, K. The chaperonin containing t-complex polypeptide 1 (TCP-1)—Multisubunit machinery assisting in protein folding and assembly in the eukaryotic cytosol. *Eur. J. Biochem.* **230**, 3–16 (1995).
- Lewis, S. A., Tian, G. L., Vainberg, I. E. & Cowan, N. J. Chaperonin-mediated folding of actin and tubulin. *J. Cell Biol.* **132**, 1–4 (1996).
- Weissman, J. S. & Kim, P. The Pro region of BPTI facilitates folding. *Cell* **71**, 841–851 (1992).
- Stewart, M. L., Grollman, A. P. & Huang, M.-T. Aurintricarboxylic acid: Inhibitor of initiation of protein synthesis. *Proc. Natl Acad. Sci. USA* **68**, 97–101 (1971).
- Frydman, J., Nimmesgern, E., Ohtsuka, K. & Hartl, F. U. Folding of nascent polypeptide chains in a high molecular mass assembly with molecular chaperones. *Nature* **370**, 111–117 (1994).

- Gesteland, R. F. Unfolding of *Escherichia coli* ribosomes by removal of magnesium. *J. Mol. Biol.* **18**, 356–371 (1966).
- Martin, J. *et al.* Chaperonin-mediated protein folding at the surface of groEL through a 'molten globule'-like intermediate. *Nature* **352**, 36–42 (1991).
- Kenney, L. J., Bauer, M. D. & Silhavy, T. J. Phosphorylation-dependent conformational changes in OmpR, an osmoregulatory DNA-binding protein of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **92**, 8866–8870 (1995).
- Martinez-Hackert, E. & Stock, A. M. The DNA-binding domain of OmpR: crystal structure of a winged helix transcription factor. *Structure* **5**, 109–124 (1997).
- Chotia, C. One thousand families for the molecular biologist. *Nature* **357**, 543–544 (1992).
- Bergman, L. W. & Kuehl, W. M. Formation of an intrachain disulfide bond on nascent immunoglobulin. *J. Biol. Chem.* **254**, 8869–8876 (1979).
- Chen, W., Helenius, J., Braakman, I. & Helenius, A. Cotranslational folding and calnexin binding during glycoprotein synthesis. *Proc. Natl Acad. Sci. USA* **92**, 6229–6233 (1995).
- Kolb, V. A., Makeyev, E. V. & Spirin, A. S. Folding of firefly luciferase during translation in a cell-free system. *EMBO J.* **13**, 3631–3637 (1994).
- Gaitanaris, G. A., Vysokanov, A., Hung, S.-Z., Gottesman, M. & Gragerov, A. Successive action of *E. coli* chaperones *in vivo*. *Mol. Microbiol.* **14**, 861–869 (1994).
- Bremer, H. & Dennis, P. P. In *Escherichia coli and Salmonella: Cellular and Molecular Biology* (ed. Neidhardt, F. C.) 1553–1569 (ASM, Washington DC, 1996).
- LaVallie, E. R. & McCoy, J. M. Gene fusion expression systems in *Escherichia coli*. *Curr. Opin. Biotechnol.* **6**, 501–506 (1995).
- Sharrocks, A. D. A T7 expression vector for producing N- and C-terminal fusion proteins with glutathione S-transferase. *Gene* **138**, 105–108 (1994).
- Hawkins, A. R. & Lamb, H. K. The molecular biology of multidomain proteins. Selected examples. *Eur. J. Biochem.* **232**, 7–18 (1995).
- Horwich, A. L., Low, K. B., Fenton, W. A., Hirschfield, I. N. & Furtak, K. Folding *in vivo* of bacterial cytoplasmic proteins: role of GroEL. *Cell* **74**, 909–917 (1993).
- Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F. & Holmes, K. C. Atomic structure of the actin: DNase I complex. *Nature* **347**, 37–44 (1990).
- Beckmann, R. P., Mizzen, L. A. & Welch, W. J. Interaction of Hsp 70 with newly synthesized proteins: implications for protein folding and assembly. *Science* **248**, 850–854 (1990).
- Nelson, R. J., Ziegelhoffer, T., Nicolet, C., Werner-Washburne, M. & Craig, E. A. The translation machinery and 70 kd heat shock protein cooperate in protein synthesis. *Cell* **71**, 97–105 (1993).
- Mayhew, M. *et al.* Protein folding in the central cavity of the GroEL–GroES chaperonin complex. *Nature* **379**, 420–426 (1996).
- Feuerstein, J., Goody, R. S. & Wittinghofer, A. Properties and characterization of nucleotide-free and metal ion-free p21 'Apoprotein'. *J. Biol. Chem.* **262**, 8455–8458 (1987).
- Pratt, J. M. In *Transcription and Translation: A Practical Approach* (eds Hames, B. D. & Higgins, S. J.) 179–210 (IRL, Oxford, 1984).
- Cameau, D. E., Ikenaka, K., Tsung, K. L. & Inouye, M. Primary characterization of the protein products of the *Escherichia coli* ompB locus: structure and regulation of synthesis of the OmpR and EnvZ proteins. *J. Bacteriol.* **164**, 578–584 (1985).
- Felner, P. L. *et al.* Lipofection: a highly efficient, lipid-mediated DNA-transfection procedure. *Proc. Natl Acad. Sci. USA* **84**, 7413–7417 (1987).
- Ausubel, F. M. *et al.* *Current Protocols in Molecular Biology* pp. 6.11.5 (Wiley, New York, 1993).
- Tucker, J. *et al.* Expression of p21 proteins in *Escherichia coli* and stereochemistry of the nucleotide-binding site. *EMBO J.* **5**, 1351–1358 (1986).

Acknowledgements. We thank K. Ewalt for assistance with the *in vivo* labelling experiments; M. K. Hayer-Hartl for the fluorescence analysis; R. Deep and E. Breslow for circular dichroism measurements; J. E. Rothman, J. Young, M. Tector and W. Houry for critically reading the manuscript; and various colleagues for providing support in the genomic analysis (listed in Fig. 1).

Correspondence and requests for materials should be addressed to F.U.H. (e-mail: uhartl@biochem.mpg.de).

KNOW YOUR COPY RIGHTS RESPECT OURS

The publication you are reading is protected by copyright law.
Photocopying copyright material without permission is no different from stealing a magazine from a newsagent, only it doesn't seem like theft.

If you take photocopies from books, magazines and periodicals at work your employer should be licensed with CLA.

Make sure you are protected by a photocopying licence.



The Copyright Licensing Agency Limited
90 Tottenham Court Road, London W1P 0LP
Telephone: 0171 436 5931 Fax: 0171 436 3986

A magnetic switch that determines the speed of astrophysical jets

D. L. Meier*, S. Edgington*†, P. Godon*, D. G. Payne‡ & K. R. Lind§

* Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California 91109, USA

† Princeton University, 67 Spelman Hall, Princeton, New Jersey 08544, USA

‡ Intel Corporation, 5200 N.E. Elam Young Parkway, Hillsboro, Oregon 97124, USA

§ Digital Equipment Corporation, LLNL, L-560, Livermore, California 94550, USA

The mechanism by which astrophysical jets form is an important factor in understanding the nature and evolution of phenomena such as active galactic nuclei and quasars, Galactic superluminal X-ray sources and young stellar objects. Of the many schemes proposed for jet production, only the magnetized accretion disk model of Blandford and Payne¹ seems to be applicable to all of these systems, and also offers the potential for generating the highly relativistic flows observed in some quasars². But the source of variation in jet morphology observed for different sources remains unclear. Here we report time-dependent numerical simulations of jet formation which show that the character and speed of the jets produced depend dramatically on whether magnetic forces dominate over gravity in the accretion disk corona. This ‘magnetic switch’ is not predicted by steady-state, self-similar disk models, or by relativistic wind theory (which generally ignores the gravitational field). The effect provides a natural explanation for the existence of two known classes of extragalactic radio source and for the variation of their properties with radio luminosity. It also provides insight into protostellar and galactic microquasar systems.

For the past several years we have been studying numerically the flows induced in the coronae of magnetized accretion disks. Our simulations begin with a situation similar to the boundary conditions used by Blandford and Payne¹ and also used in other semi-analytic studies^{3,4} (Fig. 1), but are able to investigate time-dependent and two-dimensional effects, unlike the steady-state, self-similar models. A useful parameter for measuring the importance of the magnetic field in the corona is the ratio ν of the speed of magnetic waves in the gas—the Alfvén velocity ($V_A \equiv B/(4\pi\rho)^{1/2}$, where B is the magnetic field strength and ρ is the gas density)—to the escape velocity ($V_{\text{esc}}(r) \equiv (2GM_c/r)^{1/2}$, where M_c is the mass of the central star or black hole, $r = (R^2 + Z^2)^{1/2}$ is the spherical radial coordinate and R and Z are the cylindrical radial and axial coordinates, respectively). In our studies ν can be rather large, ranging from 0.4 to 2.0 in the simulations we report here. These values are much larger than those investigated by others who have performed similar numerical simulations^{5–11}.

Most of the coronal flows that result in our simulations are decidedly jet-like, but the flow speed and character are strong functions of ν . Figure 2 shows two such cases, each with an initial field polar angle of $\theta = 54^\circ$, but with slightly different values of the Alfvén velocity in the corona. In the low-magnetic-field case (Fig. 2a; $\nu = 0.4$), the flow velocity inside the jet is of the order of the Alfvén speed— $V_{\text{jet}} \approx 0.6V_{\text{esc}} = 1.5V_A$ (as measured at $R_0 = 7.2r_c$, the place where the magnetic field attains its maximum strength, and r_c is the scale radius of the central accreting object). However, if the Alfvén velocity in the corona is increased by only a factor of three to $\nu = 1.2$ (Fig. 2b), the jet entirely changes character with a flow speed nearly 20 times that of the escape speed. We have named these two modes of flow type 1 and type 2 jets, respectively, and the transition that occurs from one type to the other we call the ‘magnetic switch’.

In addition to the simulations shown in Fig. 2, we have performed over 50 others with varying magnetic field strength and polar field angle. Those with the azimuthal field strength $B_\phi = 0$ in the corona are shown in Fig. 3, which shows that the magnetic switch is even more dramatic than Fig. 2 would indicate, with the jet velocity increasing by as much as 80 times when the Alfvén velocity is increased by only 40%. The effect has a simple physical interpretation. When $\nu < 1$ (the left-hand side of Fig. 3), gravitational forces dominate magnetic forces and the jet struggles to escape the system’s gravity. On the right side of the diagram, however, magnetic forces dominate and the system acts like a particle accelerator; the jet is not much affected by the gravity of the central object.

The behaviour of the magnetic switch is not a strong function of the polar magnetic field angle θ ; it exists for all angles studied in Fig. 3 (8° – 83°). In fact, jets still occur for $\theta < 30^\circ$ —seemingly contradicting the predictions of Blandford and Payne¹. The reason is that only in rare cases do the conditions in the disk corona match directly onto the quasi-steady-state jet solutions far above the disk. Instead, there is usually a transition region above the corona where the magnetic poloidal field and flow vectors undergo dramatic changes in direction, and the magnetic field lines dynamically acquire an angle greater than 30° . Beyond this transition region, the flow is similar in character to the solutions of Blandford and Payne¹. Thus, even disks that initially violate the $\theta > 30^\circ$ condition

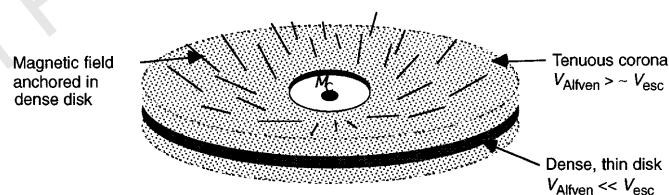


Figure 1 Schematic representation of the initial and boundary conditions of our accretion disk corona simulations. A dense accretion disk (shown dark) orbits a compact object (star or black hole) of mass M_c and scale radius r_c , which is also our unit of distance and our mesh spacing. (We take $r_c = 2 \times 10^{12}$ cm and GM_c/c^2 for the protostar and black-hole cases, respectively, so the radius at which much of the acceleration takes place ($R_0 = 7.2r_c$) is ~ 1 AU and ~ 3.6 Schwarzschild radii; see text for more discussion.) Open magnetic field lines, making an angle θ with the rotation axis, protrude from the disk and through a much less dense and more extended corona (shown stippled) whose temperature is hotter than the disk but still colder than the (virial) temperature of a halo that permeates the rest of the system. The coronal density is assumed to be ten times the asymptotic halo density; the disk density is many orders of magnitude larger than that—essentially infinite in our simulations. The corona is replenished continually from below by a wind of the same material at a velocity of only 5% of the escape speed at that point in the disk ($V_{\text{esc}} = (2GM_c/R)^{1/2}$). The sound speed in this wind is only 0.9% of the local escape speed—truly cold disk material—but, because of shock heating, the wind velocity is more representative of the eventual coronal sound speed. At their bases, the field lines are poloidal only ($B_\phi = 0$) and are anchored in the dense disk, rotating with the local keplerian angular velocity ($\Omega_K = (GM_c/R^3)^{-1/2}$) but having no appreciable radial motion. This assumption is valid for timescales short compared with the accretion drift time (~ 5 months for a $10^8 M_\odot$ active galactic nucleus (AGN) black hole, ~ 1 s for a $10 M_\odot$ Galactic black hole, and ~ 2 months for a $1 M_\odot$ protostar) and when the magnetic field is not strong enough to disrupt the disk itself ($V_A^{\text{disk}} \ll V_{\text{esc}}$). Standard accretion disk models like the α -model²⁴ naturally have a low Alfvén velocity. But if a low-density corona also exists/ V_A there can be much higher for the same magnetic field line by the ratio $(\rho_{\text{disk}}/\rho_{\text{corona}})^{1/2}$. Many accretion disks, therefore, could possess fairly benign magnetic fields that are still strong enough to appreciably affect the dynamics of their coronae ($V_A^{\text{corona}} \geq V_{\text{esc}}$). Our simulations are allowed to evolve from these initial conditions, using our axisymmetric magnetohydrodynamic (MHD) simulation code FLOW²⁵ with cylindrical radial coordinate R and axial coordinate Z . The simulation region is bounded at $Z = 0$ by the infinitely thin disk and inflow region for $5r_c \leq R \leq R_{\text{max}}$ and reflective conditions for $0 < R < 5r_c$; by the rotation axis at $R = 0$ and by open (‘outflow’) boundaries at $R = R_{\text{max}}$ and $Z = Z_{\text{max}}$.

for outflow can still form jets, making magnetohydrodynamic (MHD) jet production an even more robust mechanism than previously thought.

Our non-relativistic simulations (Fig. 3) are applicable to protostellar jets. There is, however, some uncertainty in choosing the scale radius r_c and velocity V_c , and hence, the radius $R_0 = 7.2r_c$ where much of the acceleration occurs. If, for a solar-type star, $R_0 \approx 200R_\odot = 1 \text{ AU}$, then $r_c = 2 \times 10^{12} \text{ cm}$ and $V_c = 80 \text{ km s}^{-1}$ (here $R_\odot$ is the solar radius). The magnetic switch then produces two modes of outflow. For $\nu < 1$, the outflow rate is predicted to be a few tens of km s^{-1} ; for $\nu > 1$, the rate is predicted to be several hundred km s^{-1} , which is roughly consistent with observations^{12,13}. On the other hand, if the site of acceleration is near the protostar, say $R_0 = 2R_\odot$, as some investigators have suggested, then the magnetic switch predicts a low-velocity mode of a few hundred km s^{-1} and a

much higher-velocity mode with jet velocities of several thousand km s^{-1} . The latter have not been observed, although a search for such extremely high-velocity flow, particularly in the body of known protostellar jets, would be useful. We emphasize that if a hot, tenuous corona (with $\nu > 1$) does not form near the protostar, this high velocity flow will not occur, and the magnetic switch would always be “off”.

One class of object in which two modes of jet definitely have been observed is the extragalactic radio sources, which are believed to be formed by disk accretion onto black holes¹⁴. Some radio source jets are rather slow ($V_{\text{jet}} \leq 0.6c$) and dissipate (radiating strongly at radio wavelengths) shortly after leaving the galaxy nucleus (Fanaroff and Riley (FR) class I objects^{15,16}). In contrast, other radio jets are very fast (Lorentz factor $\gamma \equiv (1 - V^2/c^2)^{-1/2} \approx 10$) and do not dissipate or radiate much until they reach large distances from the

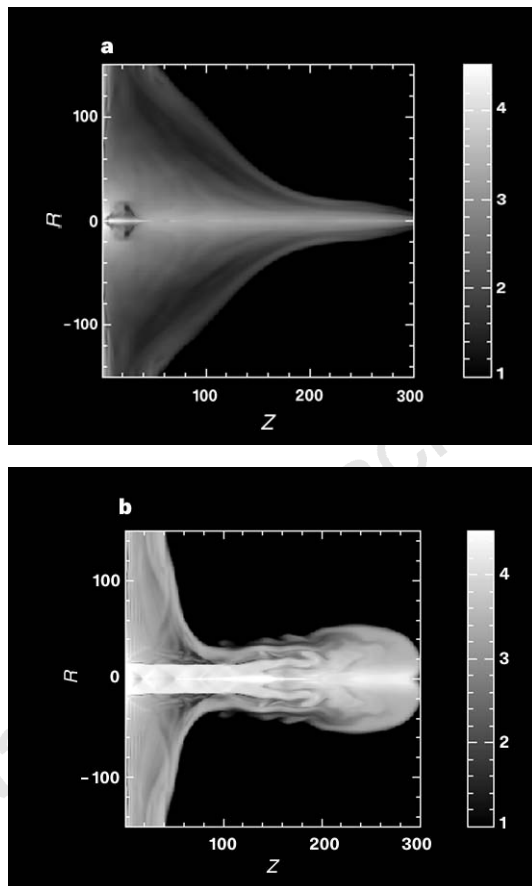


Figure 2 Logarithmic grey-scale plots of magnetic pressure ($B^2/8\pi$) for two MHD disk corona simulations that differ in only one respect: the disk coronal magnetic field strength in **b** is a factor of three larger than that in **a**. The disk in Fig. 1 lies along the left ($Z = 0$) axis. Both have a field polar angle of 54° , and we show both when the jet reaches the maximum extent of the mesh. The grey scale at the right shows relative field strength over 4.5 orders of magnitude. The simulation shown in **a** has a ratio of Alfvén velocity to escape velocity of $\nu = 0.4$ throughout the disk corona and is shown after ~ 8.6 disk rotation times (as measured at $R = R_0$). Even though the simulation begins with gravitational forces dominating, the azimuthal component of the magnetic field increases in strength owing to differential rotation, recoils upward and is able to eject a slow jet with a speed of $0.6V_{\text{esc}}(R_0)$. This is also approximately the propagation speed of the jet head through the external medium, indicating that the jet is fairly heavy. The simulation in **b** has $\nu = 1.2$ and is displayed at ~ 4.5 rotation times. This jet is produced by magneto-centrifugal acceleration rather than magnetic recoil, but still has a significant enough azimuthal field for collimation. Its internal speed is $19V_{\text{esc}}(R_0)$, over 30 times faster than the jet in **a** and ~ 20 times faster than propagation of the jet head, indicating that this jet is very light ($\rho_{\text{jet}} \ll \rho_{\text{halo}}$).

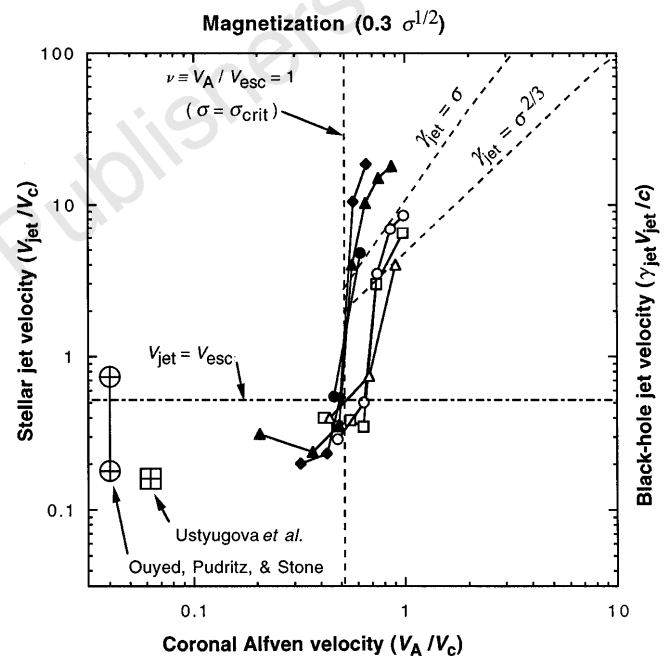


Figure 3 Jet speed as a function of coronal magnetic field strength and angle for both stellar accretion disks (lower and left axes) and black-hole disks (upper and right axes). The stellar interpretation of our multi-simulation results shows V_{jet}/V_c versus magnetic field strength in terms of $V_A(R_0)/V_c$, where V_c is the Kepler velocity at r_c ($(GM_c/r_c)^{1/2}$). We also show simulation results from ref. 10 (Ustyugova et al.) and ref. 11 (Ouyed, Pudritz and Stone). Although differing from ours in some respects, most notably in having significant gas pressure and non-zero azimuthal magnetic field strength $B_\phi(R_0) = B_{\phi 0}$ (both of which we have found tend to increase the jet velocity), their results still agree roughly with ours in the $\nu < 1$ region. The different curves show families with the same polar magnetic field angle θ : open triangles, -8° ; open squares, -24° ; open circles, -38° ; filled triangles, -54° ; filled diamonds, -68° ; filled circles, -83° . For small angles, the speed and power of the jet diminishes at the expense of a loosely- or un-collimated magnetic wind emanating from a large part of the disk outside the jet-production region. This wind also displays the magnetic switch effect. The black-hole interpretation shows the product of the jet velocity and Lorentz factor ($\gamma_{\text{jet}} \beta_{\text{jet}} = V_{\text{jet}}/[c^2 - V_{\text{jet}}^2]^{1/2}$) versus the magnetization parameter at $R_0 = 7.2GM_c/c^2$ (using $\Omega = \Omega_K$, $V_{i0} = 0.05V_{\text{esc}}$, $V_c = c$ and an injection region of $\Delta r \approx 2R_0$): $\sigma = V_A^2 V_{\text{esc}}^2 / (c^3 V_{i0}) = 11(V_A/V_c)^2$. We also plot the result from relativistic wind theory^{20,21} ($\gamma/\beta)_{\text{jet}} = \sigma^n$ for $n = 2/3$ and $n = 1$, which is valid only for $\sigma > \sigma_{\text{crit}} \approx 11(V_{\text{esc}}/c)^2 = 3.1$ for the black-hole case. Our simulations agree with relativistic wind theory in this regime to within the factor of two or so error estimated in the text. For truly general relativistic MHD simulations, the value of σ_{crit} , and the curves, will be shifted to the right by a factor of $\sim \gamma_0[1 + (V_{\text{esc}}/c)^2]^{1/2} = 1.13\gamma_0$. Although γ_0 could be as high as γ_{jet} , it is probably of the order of unity (smooth acceleration over a few R_0 in distance).

galaxy and are stopped when the flow strikes the intergalactic medium in a strong shock (FR class II). FR I jets are associated with low-power radio sources whereas FR II jets are associated with high power sources; the transition from FR I to II in radio power is especially sharp if one examines galaxies in a single optical luminosity class¹⁷.

We have found that our non-relativistic simulations can also be applied to extragalactic radio sources where relativistic flow and black-hole accretion occur. The conditions under which this is valid require $V_A \leq c$ but, surprisingly, allow jet speeds of any value, including $\gamma \gg 1$. Relativistic effects important here are an increase in mass of the gas due to kinetic and magnetic energies (which limit its velocity to less than c and increase its weight in the black-hole's gravitational field) and an electric force perpendicular to the magnetic field lines that affects only their angle¹⁸. The velocity terms in the relativistic and non-relativistic flow equations are very similar¹⁹, involving $(m_{\text{gas}} + m_{\text{mag}})u$ instead of $m_{\text{gas}}V$, where $u = \gamma V$, m_{gas} is the rest mass of the plasma, and m_{mag} is that of the magnetic field. At every point in the flow we can replace V with γV and, as long as $m_{\text{mag}} \leq m_{\text{gas}}$ (that is, $V_A \leq c$), our non-relativistic simulations can calculate γ to within a factor of two or so of the actual Lorentz factor. Field angle effects will also be within a factor of two or less in the tangent as long as $V_{\text{esc}} \leq c$; as we have investigated a range of a factor of 60 in the tangent, and observe the magnetic switch throughout, only the details of the solutions will be affected by the electric force, not our overall conclusions. Finally, the weight of the kinetic and magnetic energies will affect only the critical value of the magnetic field $B_{\text{crit}} \equiv (4\pi\rho)^{1/2}V_{\text{esc}}$ at which the magnetic switch takes place. The condition will still be $\nu = 1$, but we now must use the relativistic value of $V_{\text{esc}} = \gamma_0[2(1 + V_{\text{jet}}^2/c^2)GM_J/r]^{1/2}$ (for a Schwarzschild black hole, where γ_0 is the jet flow speed while still affected by the black hole's gravity). Although γ_0 is probably of the order of unity, general relativistic MHD simulations will be necessary to confirm this.

Special relativistic calculations of plasma flow in rotating magnetic fields have been performed extensively by many authors (refs 18–21 and references therein) in the context of pulsar and jet models. The character and speed of the flow is a function of the magnetization parameter $\sigma = V_A^2 A_0 / (2\pi R_L^2 c V_{i0})$, measured at the place where the outflow originates. (Here $R_L = c/\Omega$ is the radius of the light cylinder, Ω is the angular velocity of the rotating magnetic field, A_0 is the disk surface area over which outflow takes place and V_{i0} is the injection velocity—equivalent to our injected coronal wind velocity.) The terminal velocity of the flow is given by the simple expression $\gamma_{\text{jet}} V_{\text{jet}}/c \approx \sigma^n$, where n lies between 2/3 and 1, depending on exactly how the flow collimates^{20,21}. We have plotted these curves in Fig. 3, with σ on the upper axis, and show that the theory agrees well with our results for $\nu > 1$. However, this relation was derived under the assumption that the gravitational force is negligible. Current relativistic wind theory, therefore, is not valid in the region of parameter space which the magnetic switch occurs. Indeed, we can define a critical value of the magnetization ($\sigma_{\text{crit}} \equiv V_{\text{esc}}^2 A_0 / (2\pi R_L^2 c V_{i0})$), which is of the order of unity for the black-hole case) below which relativistic wind theory breaks down because gravitational forces become important. For $\sigma < \sigma_{\text{crit}}$ ($\nu < 1$), the jet velocity drops rapidly from the $\gamma \geq 10$ result of relativistic wind theory to $V_{\text{jet}} \approx (0.1–0.4)c$.

The magnetic switch occurs in velocity only, not in total jet power. Whereas the jet velocity changes dramatically as the strength of the field is increased through the critical value, we find that the total jet power varies smoothly with the square of the magnetic field, as predicted by Blandford and Payne¹, $P_{\text{jet}} \approx B_0^2 R_0^2 (GM_J/R_0)^{1/2}$. For $\nu < 1$, most of the power is in advected magnetic field, whereas for $\nu > 1$ the jet carries a significant amount of kinetic power. Both types of power are available for accelerating particles and producing synchrotron radio emission.

The magnetic switch provides a natural explanation for the

existence of the FR I and II classes of radio source, with the former being produced by accretion disks with $\nu < 1$ and the latter for $\nu > 1$. Predicted jet speeds agree with observations in each case ($V_{\text{jet}} \leq 0.6c$ for FR I, $\gamma \approx 10$ or higher for FR II). In addition, the effect explains why the two classes are associated with low- and high-power radio sources, respectively, and why the transition is so sharp: both jet power and jet speed are direct functions of the strength of the magnetic field in the disk corona, but whereas the former is a smooth function the latter is a strong one. Additional results on variation of the FR I/FR II break with galaxy optical luminosity will be reported elsewhere (D.L.M., manuscript in preparation).

Because our model states that jets in sources with FR I morphology are formed in the nucleus with sub- or trans-relativistic velocities, it predicts that, statistically, observations of FR I sources by Very Long Baseline Interferometry will show primarily low jet velocities ($V_{\text{jet}} < 0.6c$), whereas FR IIs should have much faster VLBI jets. A corollary is that the parent objects of most BL Lac sources, which have $\gamma = 2–5$, cannot have FR I morphology.

Bimodal jet ejection has also been observed in the galactic microquasars GRO J1655–40 and Cygnus X-3 (ref. 22). Magnetic switching can also explain such behaviour, but a detailed determination of the bolometric luminosity of these sources is needed in order to determine the exact triggering mechanism. Possible choices are (1) very rapid (“super-Eddington”) accretion that smothers the magnetic field²³ or (2) a modest increase in the (more normal, “sub-Eddington”) accretion rate that can cool and destroy a low-density corona with a large flux of soft photons. $\square$

Received 6 January; accepted 3 June 1997.

- Blandford, R. D. & Payne, D. G. Hydromagnetic flows from accretion discs and the production of radio jets. *Mon. Not. R. Astron. Soc.* **199**, 883–903 (1982).
- Vermeulen, R. C. in *Extragalactic Radio Sources* (eds Ekers, R., Fanti, C. & Padrielli, L.) 56–62 (IAU Symp. No. 175, Kluwer, Dordrecht, 1996).
- Lovelace, R. V. E., Mehanian, C., Mobarri, C. M. & Sulkanen, M. E. Theory of axisymmetric magnetohydrodynamic flows—disks. *Astrophys. J. Suppl. Ser.* **62**, 1–37 (1986).
- Kudoh, T. & Shibata, K. Mass flux and terminal velocities of magnetically driven jets from accretion disks. *Astrophys. J.* **452**, L41–L44 (1996).
- Uchida, Y. & Shibata, K. Magnetodynamical acceleration of CO and optical bipolar flows from the region of star formation. *Publ. Astron. Soc. Jpn.* **37**, 515–535 (1985).
- Stone, J. M. & Norman, M. L. Numerical simulations of magnetic accretion disks. *Astrophys. J.* **433**, 746–756 (1994).
- Bell, A. R. & Lucsek, S. G. Magnetohydrodynamic jet formation. *Mon. Not. R. Astron. Soc.* **277**, 1327–1340 (1995).
- Matsumoto R. in *Accretion Phenomena and Related Outflows* (IAU Colloq. No. 163, in the press).
- Koide, S., Shibata, K. & Kudoh, T. in *Accretion Phenomena and Related Outflows* (IAU Colloq. No. 163, in the press).
- Ustyugova, G. V., Koldoba, M. M., Romanova, V. M., Chechekin, V. M. & Lovelace, R. V. E. Magnetohydrodynamic simulations of outflows from accretion disks. *Astrophys. J.* **439**, L39–L42 (1995).
- Ouyed, R., Pudritz, R. E. & Stone, J. M. Episodic jets from black holes and protostars. *Nature* **385**, 409–414 (1997).
- Bachiller, R. Bipolar molecular outflows from young stars and protostars. *Annu. Rev. Astron. Astrophys.* **34**, 111–154 (1996).
- Eisloffel, J. in *Jets from Stars and Galactic Nuclei* (ed. Kundt, W.) 104–117 (Lecture Notes in Physics, Vol. 471, Springer, Berlin, 1996).
- Blandford, R. D. & Rees, M. J. in *Testing the AGN Paradigm* (eds Holt, S. S., Neff, S. G. & Urry, C. M.) 3–19 (AIP Conf. No. 254, Am. Inst. Phys., College Park, MD, 1991).
- Fanaroff, B. L. & Riley, J. M. The morphology of extragalactic radio sources of high and low luminosity. *Mon. Not. R. Astron. Soc.* **167**, 31P–35P (1974).
- Bicknell, G. V. The relationship of the Fanaroff–Riley classification of extragalactic radio-sources to jet physics. *Publ. Astron. Soc. Aust.* **6**, 130–137 (1985).
- Ledlow, M. J. & Owen, F. N. 20-cm VLA survey of Abell clusters of galaxies. VI. Radio/optical luminosity functions. *Astron. J.* **112**, 9–22 (1996).
- Li, Z.-Y., Chieue, T. & Begelman, M. C. Electromagnetically driven relativistic jets: a class of self-similar solutions. *Astrophys. J.* **394**, 459–471 (1992).
- Goldreich, P. & Julian, W. H. Stellar winds. *Astrophys. J.* **160**, 971–977 (1970).
- Henriksen, R. N. in *Accretion Disks and Magnetic Fields in Astrophysics* (ed. Bevilacqua, G.) 117–128 (Kluwer, Dordrecht, 1989).
- Camenzind, M. in *Accretion Disks and Magnetic Fields in Astrophysics* (ed. Bevilacqua, G.) 129–143 (Kluwer, Dordrecht, 1989).
- Hjellming, R. M. in *Accretion Phenomena and Related Outflows* (IAU Colloq. No. 163, in the press).
- Meier, D. L. in *Accretion Phenomena and Related Outflows* (IAU Colloq. No. 163, in the press).
- Shakura, N. I. & Sunyaev, R. A. Black holes in binary systems. Observational appearance. *Astron. Astrophys.* **24**, 337–355 (1973).
- Lind, K. R., Payne, D. G., Meier, D. L. & Blandford, R. D. Numerical simulations of magnetized jets. *Astrophys. J.* **344**, 89–103 (1989).

Acknowledgements. We thank R. Linfield, R. Ouyed, R. Hjellming and K. Meisenheimer for comments on the manuscript. Part of this research was carried out at the Jet Propulsion Laboratory, under contract to NASA. S.E. was an associated Western Universities Summer fellow; P.G. was a National Research Council research associate at the NASA Jet Propulsion Laboratory.

Correspondence should be addressed to D.L.M. (e-mail: dlm@cena.jpl.nasa.gov).

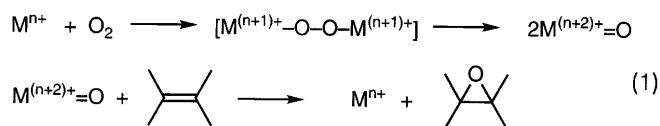
A ruthenium-substituted polyoxometalate as an inorganic dioxygenase for activation of molecular oxygen

Ronny Neumann & Mazal Dahan

Casali Institute of Applied Chemistry, Graduate School of Applied Science,
The Hebrew University of Jerusalem, Jerusalem, Israel 91904

The development of a procedure for the epoxidation of alkenes with molecular oxygen is an important industrial and synthetic goal. Non-radical activation of dioxygen by monooxygenase enzymes such as the iron-porphyrin-based cytochrome P-450 involves the insertion of one oxygen atom of O₂ into the organic substrate while the other oxygen is reduced to water in the presence of an electron donor. Dioxygenase enzymes, on the other hand, catalyse insertion of both oxygen atoms without reducing agents. Oxidation of hydrocarbons with dioxygen catalysed by transition-metal compounds invariably proceeds by free-radical pathways rather than by dioxygenase-type reactions, and so do not allow the epoxidation of alkenes with molecular oxygen. The sterically hindered ruthenium tetramesityl porphyrin complex has been shown previously to activate dioxygen in a dioxygenase mode¹. We describe here the use of the polyoxometalate {[WZnRu₂(OH)(H₂O)](ZnW₉O₃₄)₂}¹¹⁻ (ref. 2) as a catalyst for non-radical molecular oxygen activation and alkene epoxidation. The polyoxometalate can be considered to act as an inorganic dioxygenase catalyst. The advantages of using inorganic catalysts such as polyoxometalates, as opposed to organometallic complexes, is their well documented stability against decomposition by self-oxidation^{3,4}.

The epoxidation of alkenes is a reaction of significant synthetic importance. Classically, epoxidation is carried out using organic peracids as epoxidizing reagents, or by intramolecular etherification of intermediate chlorohydrins formed by reaction of alkenes with hypochlorous acid. Safety and environmental concerns have led to alternative schemes, most notably catalytic epoxidation with alkyl hydroperoxides and hydrogen peroxide. On the other hand, transition-metal-catalysed oxidation of alkenes with molecular oxygen leads primarily to allylic oxidation products via auto-oxidation pathways rather than epoxides⁵. Catalytic oxygen transfer from molecular oxygen to alkenes to form epoxides without use of sacrificial reducing agents is therefore an important goal. One simplified catalytic approach to such an oxygen transfer is outlined in equation (1).



Our previous study⁶ into molecular oxygen activation in the hydroxylation of adamantane with the ruthenium-substituted polyoxometalate {[WZnRu₂(OH)(H₂O)](ZnW₉O₃₄)₂}¹¹⁻ (1, Fig. 1) showed three lines of preliminary evidence that the reaction was not of a free-radical nature. (1) Hydroxylation was exclusively at the tertiary carbon as compared to normal free-radical oxidation which gives a tertiary:secondary ratio of 3:1; (2) radical scavengers had no effect on catalytic activity; and (3) substrates very susceptible to auto-oxidation such as cumene were much less reactive compared to adamantane. These results led to further research into the mechanism of the molecular oxygen activation in the hydroxylation

of adamantane. First, we have consolidated the evidence for a non-free-radical pathway. Hydroxylation of adamantane-1,3-*d*₂ (97.2% purity by mass spectrometry) under standard conditions (0.25 mmol substrate, 0.25 μmol Q₁₁-1 (where Q is tricaprylmethyl ammonium), 0.5 ml 1,2-dichloroethane (previously treated with SnCl₂ and H₂SO₄ to remove peroxides and stabilizers, respectively, followed by distillation), 10 ml O₂ at 1 atm, 80 °C, 24 hours) yielded an isotope ratio (1-adamantan-3,5-*d*₂/1-adamantan-3-*d*₁) of 5.7 ± 0.2 (ref. 7) as compared to values of 2.0 ± 0.3 found for free-radical reactions using a simple PCoW₁₁O₃₉⁵⁻ polyoxometalate as catalyst. A reaction profile for the adamantane hydroxylation (Fig. 2) showed a long induction period of 10 hours before initiation of the reaction. Investigation of the source of the induction period was carried out as follows. Adamantane and 1 in 1,2-dichloroethane were heated under argon for 12 hours at 80 °C; oxygen was then added and the reaction continued for 3 hours at 80 °C. No hydroxylation was observed. On the other hand, similar heating of 1 in 1,2-dichloroethane under 1 atm oxygen for 12 hours, followed by

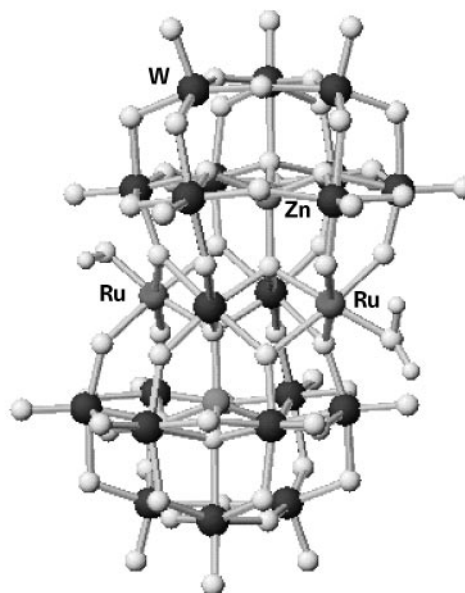


Figure 1 The {[WZnRu₂(OH)(H₂O)](ZnW₉O₃₄)₂}¹¹⁻ polyoxometalate; species 1.

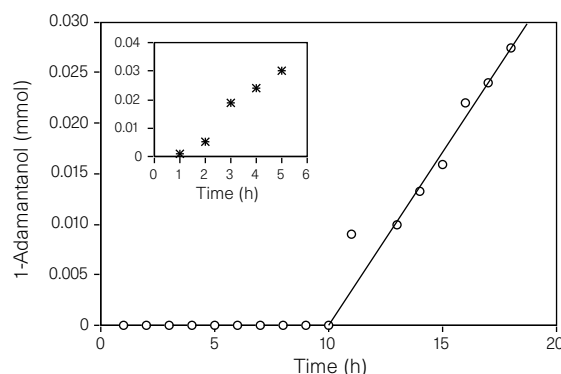


Figure 2 A reaction profile for adamantane hydroxylation. (Reaction conditions: 0.25 mmol adamantane, 0.25 μmol Q₁₁-1, 0.5 ml 1,2-dichloroethane, 10 ml O₂ at 1 atm, 80 °C, 24 h). The induction period was a function of temperature (~8 h at 100 °C; ~6 h at 120 °C). The rate as a function of temperature (5 points, *r*² = 0.96) gave Δ*H*₂₅[‡] = 7.22 kcal mol⁻¹ and Δ*S*₂₅[‡] = -54.9 cal mol⁻¹ K⁻¹, where Δ*H*[‡] and Δ*S*[‡] are respectively the enthalpy and entropy of activation. Inset, the reaction profile obtained by addition of ~0.5 μmol Zn metal before initiating the reaction by pressurizing with 1 atm O₂.

addition of adamantane and further heating for 3 hours, gave 4 mol% 1-adamantanol (40 turnovers). It thus became clear that the induction period is related to the activation of the ruthenium polyoxometalate with molecular oxygen.

The reaction of **Q**₁₁-**1** with dioxygen was further studied by infrared spectroscopy (Fig. 3). We treated 5 μ mol **Q**₁₁-**1** in 0.5 ml 1,2-dichloroethane with 1 atm oxygen at 80 °C for 12 hours. The three-peak spectrum of untreated **Q**₁₁-**1**, trace c in Fig. 3, has the following features: W–O (terminal) 928 cm^{−1}; W–O–W (corner-sharing octahedra) 881 cm^{−1}; W–O–W (edge-sharing octahedra) 765 cm^{−1}. On treatment with ¹⁶O₂ a new peak at 805 cm^{−1} (see trace b in Fig. 3) is observed. This vibration may be attributed either to a bridging peroxo species or, in analogy to the ruthenium porphyrin system, to a ruthenium-oxo species. Similar treatment with ¹⁸O₂ yields a similar peak but shifted to lower energy at 720 cm^{−1} (trace a in Fig. 3). This shift of 85 cm^{−1} is somewhat greater than the expected shift of 40 cm^{−1}. It should be noted that it is not unusual for infrared isotope shifts for M=O to differ from the expected value for metal-oxo absorptions. For example, in certain ruthenium-oxo⁸ and other metal oxo⁹ species, unexpected shifts have been observed. Other research has shown a large influence of the remaining ligation sphere on the magnitude of the infrared isotope shift¹⁰. Consecutive treatment of **Q**₁₁-**1** with ¹⁶O₂, vacuum then Ar (three times) and ¹⁸O₂ and vice versa showed that the formation of the infrared peaks were reversible, to give the original spectrum. Evidence for the interaction of the ruthenium polyoxometalate with oxygen was also found in the ultraviolet–visible spectrum. Thus, **Q**₁₁-**1** in 1,2-dichloroethane absorbs strongly at 262 nm (log ϵ = 4.70, where ϵ is molar absorptivity). On treatment with O₂ as described above, a distinct peak at 290 nm (log ϵ = 4.55) was formed. This peak is attributable to a ligand-to-metal charge transfer transition, although an admixture of such charge transfer and *d*–*d* transitions can also be contemplated. Electron-spin resonance (ESR) spectroscopy showed that the proposed intermediate ‘ruthenium–oxygen’ species was ESR-silent; ¹⁸³W NMR was impossible owing to insufficient sensitivity at feasible concentrations.

After treatment with dioxygen for 12 hours at 80 °C the ruthenium-substituted polyoxometalate was frozen and excess oxygen was removed by vacuum. A stoichiometric amount of alkene was added and the solution kept at room temperature

(~22 °C) for 4 hours. From the results in Table 1, one clearly observed the selective and high-yield formation of epoxides due to the reaction with the active ‘ruthenium–oxygen’ species. The formation of the active species occurred without formation of water. This was shown by catalytic adamantane hydroxylation using ¹⁸O₂ (98% purity). After 100 turnovers, dried molecular sieves with a pore size of 4 Å were added to the reaction mixture and filtered off. Under water desorption conditions, 270 °C, the evolving gas from the molecular sieves was directly sampled by mass spectroscopy. No H₂¹⁸O was observed. Interestingly, catalytic epoxidation under standard catalytic reaction conditions (that is, mixing alkene, **Q**₁₁-**1** and dioxygen in solvent) yielded no reaction. We consider that in the presence of alkene, the complexation of dioxygen to the ruthenium active centre is totally inhibited by alkene coordination. Indeed, attempts to carry out adamantane hydroxylation in the presence of an equivalent amount of cyclooctene yielded neither adamantane hydroxylation nor cyclooctene epoxidation reactions.

The above results show the formation of a ‘ruthenium–oxygen’ intermediate species without co-formation of water and demonstrate its ability to promote epoxidation of alkenes. Above we have presented a large body of evidence showing that **1** catalyses oxygenation in a dioxygenase mode, although some contribution of free-radical chemistry is still a remote possibility. There remained two questions. First, to gain an understanding of why {[WZnRu₂(OH)(H₂O)](ZnW₉O₃₄)₂}^{11−} is unique in its oxygenase activity compared to other ruthenium-containing polyoxometalates. Second, to give a more specific identification of the active species. Considering the first point, we have shown previously that ruthenium-substituted Keggin compounds give a free-radical reaction profile in alkane oxidation¹¹. Another ruthenium-containing Dawson–Wells polyoxometalate has yielded an inactive ruthenium μ -oxo dimer in the presence of dioxygen¹². Modelling of the interaction of dioxygen with **1** using the Cache molecular modelling program shows that the ruthenium atom is positioned in a sterically hindered pocket within the polyoxometalate framework. In the other ruthenium-containing polyoxometalates there is no such pocket. This placement prevents the formation of an inactive μ -oxo dimer and allows formation of a μ -peroxo intermediate. This is analogous to what has been observed in the ruthenium porphyrin system¹, where the sterically hindered tetramesityl derivative is

Figure 3 The infrared spectra of **1** before and after activation with molecular oxygen. Spectra were measured by placing a 0.02 M solution of **Q**₁₁-**1** on a KBr plate and evaporating the 1,2-dichloroethane solvent with a gentle stream of air. Trace c, untreated **Q**₁₁-**1**; trace b, **Q**₁₁-**1** treated with ¹⁶O₂ at 80 °C for 12 h; trace a, **Q**₁₁-**1** treated with ¹⁸O₂ at 80 °C for 12 h.

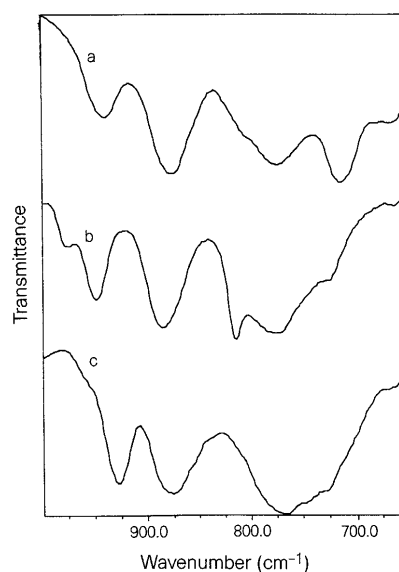


Table 1 Epoxidation of alkenes with molecular oxygen activated by **1**

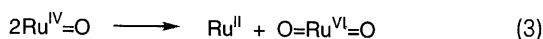
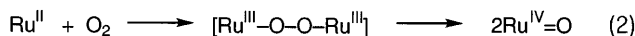
Alkene	Yield (mol%)	Selectivity (mol%)
Cyclooctene	67.5	(>99)
Norbornene	82.1	(>99)
2,3-dimethyl-2-butene	78.4	(>99)
1-octene	17.2	(>99)
Propene*	Unknown	(95)

1 is {[WZnRu₂(OH)(H₂O)](ZnW₉O₃₄)₂}^{11−}; Fig. 1. Reaction conditions: a 5 ml solution of 1,2-dichloroethane containing 100 μ mol **Q**₁₁-**1** was heated at 80 °C for 12 h under 1 atm molecular O₂ in a closed vessel. The solution was frozen, oxygen removed, and brought to room temperature. Substrate (100 μ mol) was added and the solution was left for two 2 h at room temperature and then analysed by gas-liquid chromatography.

* Propene was added at 1 atm; the conversion could not be measured. The co-product (~5 mol%) was acrolein.

catalytically active, but the non-hindered tetraphenyl species is inactive.

As noted above, the intermediate active species is ESR-silent. This seems to indicate a diamagnetic ruthenium-oxo species (for example, $\text{Ru}^{\text{IV}}=\text{O}$ or $\text{O}=\text{Ru}^{\text{VI}}=\text{O}$) or a coupled paramagnetic species ($\text{Ru}^{\text{III}}-\text{O}-\text{O}-\text{Ru}^{\text{III}}$) as the active intermediate. These conjectures are supported by our infrared spectra and by analogy with the porphyrin system¹. A ruthenium(v) active species may be ruled out because an ESR-active intermediate is formed in the presence of periodate¹³. As the $\{[\text{WZnRu}^{\text{III}}(\text{OH})(\text{H}_2\text{O})](\text{ZnW}_9\text{O}_{34})_2\}^{11-}$ compound is only catalytically active after a long induction period, we propose slow formation of a ruthenium(IV)-oxo species, followed by fast reaction under turnover conditions. We found that the induction period could be eliminated by the *in situ* reduction of the original Ru^{III} compound to the Ru^{II} analogue with two equivalents of zinc metal under argon (Fig. 2 inset). Addition of dioxygen immediately brought about adamantane hydroxylation at a rate slightly higher than the expected post-induction rate. We therefore propose that dioxygen activation, as shown in equation (2) below, occurs. This may or may not be followed by a disproportionation reaction, equation (3), as found in the ruthenium porphyrin system.



Any of the oxygen-containing intermediates shown could be the active species. We believe, however, that it is more likely that the active species is a $\text{Ru}^{\text{IV}}=\text{O}$ intermediate, which has been shown to be active for epoxidation in sterically crowded ruthenium complexes¹⁴. *Trans*-ruthenium dioxo species, active in epoxidation, seem to be unlikely, given the crowded tetragonal coordination of ruthenium in the polyoxometalate in contrast to the advantageous square planar coordination of ruthenium in the porphyrin¹. Conceivable *cis*-ruthenium dioxo species are also unlikely, as they have not been shown to promote epoxidation.¹⁵ □

Received 19 November 1996; accepted 28 May 1997.

- Groves, J. T. & Quinn, R. Aerobic epoxidation of olefins with ruthenium porphyrin catalysts. *J. Am. Chem. Soc.* **107**, 5790–5792 (1985).
- Neumann, R. & Khenkin, A. M. Noble metal (Ru^{III} , Pd^{II} , Pt^{II}) substituted “sandwich” type polyoxometalates: preparation, characterization and catalytic activity in oxidations of alkanes and alkenes by peroxides. *Inorg. Chem.* **34**, 5733–5760 (1995).
- Hill, C. L. & Prosser-McCarthy Homogeneous catalysis by transition metal oxygen anion clusters. *Coord. Chem. Rev.* **143**, 407–455 (1995).
- Okuhara, T., Mizuno, N. & Masons, M. Catalytic properties of heteropoly compounds. *Adv. Catal.* **41**, 113–252 (1996).
- Sheldon, R. A. & Kochi, J. K. *Metal Catalyzed Oxidations of Organic Compounds* (Academic, New York, 1981).
- Neumann, R., Khenkin, A. M. & Dahan, M. Hydroxylation of alkanes with molecular oxygen catalysed by a new ruthenium substituted polyoxometalate, $[\text{WZnRu}_2^{\text{III}}(\text{OH})(\text{H}_2\text{O})(\text{ZnW}_9\text{O}_{34})_2]^{11-}$. *Angew. Chem. Int. Edn. Engl.* **34**, 1587–1589 (1995).
- Sorokin, A., Robert, A. & Meunier, B. Inter-molecular kinetic isotope effect in alkane hydroxylations catalyzed by manganese and iron porphyrin complexes. *J. Am. Chem. Soc.* **115**, 7293–7299 (1993).
- Seok, W. K. & Meyer, T. J. Multiple electron oxidation of phenols by an oxo-complex of ruthenium(IV). *J. Am. Chem. Soc.* **110**, 7358–7367 (1988).
- Collman, J. P. *et al.* A synthetic analogue for the oxygen binding site in cytochrome C oxidase. *J. Am. Chem. Soc.* **116**, 9783–9784 (1994).
- Paeng, I. R. & Nakamoto, K. Resonance Raman spectra of reaction intermediates in the oxidation process of ruthenium(II) and iron(II) porphyrins. *J. Am. Chem. Soc.* **112**, 3289–3297 (1990).
- Neumann, R. & Dahan, M. Transition metal substituted Keggin type polyoxomolybdates as bifunctional catalysts for the epoxidation of alkenes by molecular oxygen. *J. Chem. Soc. Chem. Commun.* 171–172 (1995).
- Randell, W. J., Weakley, T. J. R. & Finke, R. G. Oxidation resistant inorganic porphyrin analog polyoxometalates. 3. The synthesis and X-ray crystallographic characterization of a new heteropolyoxoanion structural type, the diruthenium-oxo-bridged “bimetallic unorganic-porphyrin analog $\text{KLi}_{15}[\text{O}(\text{Ru}^{\text{IV}}\text{Cl}(\alpha_2\text{-P}_2\text{W}_6\text{O}_{61}))_2]\cdot 2\text{KCl}\cdot 60\text{H}_2\text{O}$ ”. *Inorg. Chem.* **32**, 1068–1071 (1993).
- Neumann, R. & Abu-Gnim, C. Alkene oxidation catalysed by a ruthenium substituted heteropolyanion, $\text{SiRu}(\text{H}_2\text{O})\text{W}_{11}\text{O}_{39}$: the mechanism of the periodate mediated oxidative cleavage. *J. Am. Chem. Soc.* **112**, 6025–6031 (1990).
- Bailey, C. & Drago, R. S. Catalytic epoxidation of alkenes by sterically crowded ruthenium complexes. *J. Chem. Soc. Chem. Commun.* 179–180 (1987).
- Goldstein, A. A., Beer, R. H. & Drago, R. S. Catalytic oxidation of hydrocarbons with O_2 or H_2O_2 using a sterically hindered ruthenium complex. *J. Am. Chem. Soc.* **116**, 2424–2429 (1994).

Acknowledgements. We thank A. Sorokin for the sample of 1,3-adamantane-*d*₂. This work was supported by the Basic Research Foundation administered by the Israel Academy of Sciences and Humanities.

On/off blinking and switching behaviour of single molecules of green fluorescent protein

Robert M. Dickson*, Andrew B. Cubitt†, Roger Y. Tsien‡ & W. E. Moerner*

* Department of Chemistry and Biochemistry 0340, University of California San Diego, La Jolla, California 92093-0340, USA

† Aurora Biosciences, 11149 North Torrey Pines Road, La Jolla, California 92037, USA

‡ Department of Pharmacology and Howard Hughes Medical Institute 0647, University of California San Diego, La Jolla, California 92093-0647, USA

Optical studies of individual molecules at low and room temperature can provide information about the dynamics of local environments in solids, liquids and biological systems unobscured by ensemble averaging^{1–14}. Here we present a study of the photo-physical behaviour of single molecules of the green fluorescent protein (GFP) derived from the jellyfish *Aequorea victoria*. Wild-type GFP and its mutant have attracted interest as fluorescent optical labels because the fluorophore may be formed *in vivo*^{15,16}. GFP mutants immobilized in aerated aqueous polymer gels and excited by 488-nm light undergo repeated cycles of fluorescent emission (‘blinking’) on a timescale of several seconds—behaviour that would be unobservable in bulk studies. Eventually the individual GFP molecules reach a long-lasting dark state, from which they can be switched back to the original emissive state by irradiation at 405 nm. This suggests the possibility of using these GFPs as fluorescent markers for time-dependent cell processes, and as molecular photonic switches or optical storage elements, addressable on the single-molecule level.

Wild-type GFP has two absorption maxima—a strong peak at 396 nm, and a weak one at 476 nm (ref. 15)—but only one emission peak at 508 nm, indicating that excited-state relaxation to a common emissive level occurs. The absorption peaks correspond to the neutral and anionic chromophore states, respectively, which can be interconverted by an excited-state proton transfer and spontaneous ground-state processes^{16–18}. Ser 65 (serine, S) and Thr 203 (threonine, T) are particularly close to the chromophore^{19,20} and strongly influence its photophysical properties. Alteration of Ser 65 strongly favours the anionic form (which has a long-wavelength absorption), producing mutants which can easily be pumped with readily available green lasers^{16,19,20}. Aromatic residues phenylalanine (F) and tyrosine (Y) at position 203 further increase the peak excitation wavelength by 13–24 nm, probably by increasing the polarizability around the chromophore through $\pi-\pi$ interactions¹⁹. Because of the higher brightness compared to wild type, we focus here on two particular yellow-fluorescing GFP mutants, S65G/S72A/T203F (denoted T203F) and S65G/S72A/T203Y (denoted T203Y), which differ only by the presence of a hydroxyl group.

These mutants’ long-wavelength absorption and emission properties (Fig. 1) enable facile excitation with an argon-ion laser, while allowing high-efficiency fluorescence collection and background rejection. In the bulk, T203F (Fig. 1a) has a dominant absorption at 510 nm, and a weak peak near 405 nm. Excitation at 488 nm produces long-wavelength emission, whereas excitation at 400 nm gives emission predominantly peaking at 455 nm rather than 525 nm (Fig. 1a inset). By analogy to previously studied mutants, the long-wavelength absorption and emission probably correspond to the anionic form, and the short-wavelength absorption probably arises from a small fraction of molecules with neutral chromophores^{16,17}. In contrast to T203F, T203Y (Fig. 1b) shows

much weaker absorption at 400 nm which produces (Fig. 1b inset) only weak emission at 470 nm and dominant long-wavelength fluorescence (535 nm). Excitation at 488 nm, however, yields emission nearly identical to that of T203F. The Ser65 and Thr203 mutations combine to make the low-energy absorption strongest in both mutants (opposite to wild-type photophysics); however, the emission properties suggest that a larger portion of the neutral form persists in T203F.

Single molecules of T203F (number of molecules, $n = 178$) and T203Y ($n = 199$) immobilized in polyacrylamide gels²¹ were directly imaged by excitation at 488 nm and detection of polarization-dependent emission. During several minutes of illumination, each molecule produced several seconds of fluorescence, then several seconds without emission, followed by resumption of emission, repeating over the course of many minutes. Such high-contrast 'blinking' behaviour, which is unobservable in bulk experi-

ments, is reminiscent of spectral and amplitude fluctuations reported previously for single small fluorophores^{9,10}. These GFP mutants, however, show further interesting characteristics. Each GFP molecule generally ended in a non-fluorescent state after emission of $\sim 10^6$ photons. This final dark state was found to be long-lived: after 5 min with no pumping, the single molecules remained non-emissive on re-irradiation at 488 nm (Fig. 2, panels 1, 3, 5, 7, 9). The bright state could be reproducibly recovered, however, by irradiation at 405 nm for 5 min before 488-nm re-irradiation (Fig. 2, panels 2, 4, 6, 8, 10), an optically induced switching effect. To our knowledge, these GFP mutants provide the first example of a room-temperature optical switch in which each molecule is individually addressable, and not being accomplished quasi-non-destructively by way of fluorescence (similar to earlier liquid-helium-temperature systems²²). The bacteriorhodopsin protein also exhibits room-temperature switching, but it

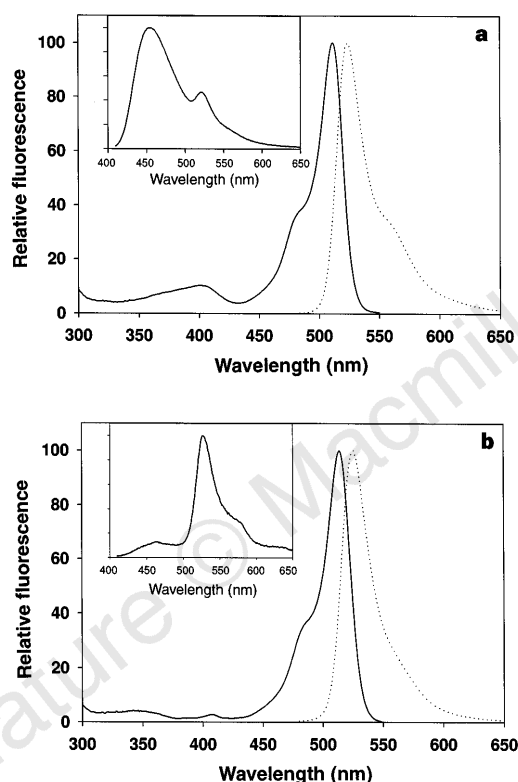


Figure 1 Bulk spectra of GFP mutants. **a**, T203F. Main figure: solid line, absorption; dotted line, emission excited by 488-nm irradiation. Inset, emission excited by 400-nm irradiation. **b**, T203Y. Main figure: solid line, absorption; dotted line, emission excited by 488-nm irradiation. Inset, emission excited by 400-nm irradiation.

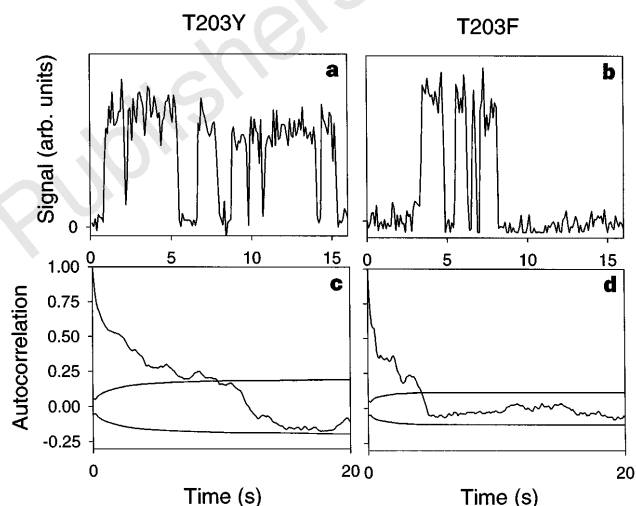


Figure 3 Traces of typical fluorescence intensity versus time for T203Y (**a**) and T203F (**b**) with an excitation intensity of $2,000 \text{ W cm}^{-2}$. Each trace covers 16 s. Autocorrelations of the fluorescence trajectories **a** and **b**, expanded to illustrate the decay over the first 20 s, are shown in **c** and **d**, respectively, with confidence limits.

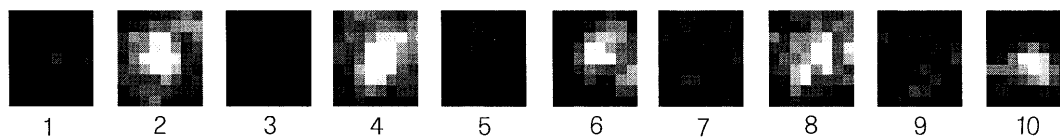


Figure 2 The switching behaviour of a single T203F molecule, illustrated by a series of ten consecutive experiments on the same molecule. Each 100-ms frame shows the emission produced by 488-nm illumination (60 nm $\times$ 60 nm pixel size). In odd-numbered frames (1, 3, 5, 7, 9), the sample was pre-illuminated for 90 s at 488 nm to prepare the long-lived dark state, followed by 5 min with no light incident, after which the displayed image was obtained. The dark state was still present in each case. In even-numbered frames (2, 4, 6, 8, 10), the sample was also

pre-illuminated for 90 s at 488 nm to prepare the dark state, but then the sample was irradiated for 5 min at 405 nm (1 W cm^{-2}), after which the displayed image was immediately obtained. In each case, the emissive state is restored by the 405-nm irradiation. The odd-numbered cases assess the thermal stability of the dark state, whereas the even-numbered cases determine its photosensitivity to 405-nm irradiation. Similar optical switching behaviour occurred for T203Y.

efficiently photoisomerizes without fluorescing²³, rendering it inaccessible to single-molecule detection and use.

Many single GFP molecule fluorescence trajectories (Fig. 3a, b) were analysed to develop on-time and off-time histograms commonly used in single-ion-channel recordings²⁴. Only the high-intensity data ($2,000 \text{ W cm}^{-2}$), however, produced a sufficient signal-to-noise ratio to reliably distinguish bright and dark by a single threshold. The on-time histograms were fitted by single exponentials (time constants: T203Y 0.86 s ($n = 89$), T203F 0.71 s ($n = 97$)). The dark-time histograms, however, showed bi-exponential decays for both mutants with a short decay time of $\sim 1 \text{ s}$ and a long decay time greater than several tens of seconds (limited by our 90-s data collection time), corresponding to the blinking and switching timescales, respectively.

To obtain more information from fluorescence trajectories at lower intensity, we use autocorrelation analysis, which has helped the understanding of several single-molecule systems¹⁴. Typical T203Y and T203F single-molecule autocorrelations are presented with confidence limits in Fig. 3c, d. The autocorrelation time τ_c (determined from an exponential fit to the autocorrelation decay) reflects the average period of on/off behaviour. Histograms of the τ_c values from 377 individual GFP molecules are shown in Fig. 4 as a function of excitation intensity. Whereas the average autocorrelation time, τ_c^{avg} , is reduced by only a factor of two for a 20-fold increase in power, the single-molecule measurements produce the entire distribution, providing clear evidence that τ_c shortens with increased intensity. As the autocorrelation measures how the fluorescence intensity correlates with itself in time, the faster and less power-dependent autocorrelation decays of T203Y relative to those of T203F indicate that the additional hydroxyl group of T203Y facilitates a lower-energy pathway to chromophore activation/inactivation (Fig 5).

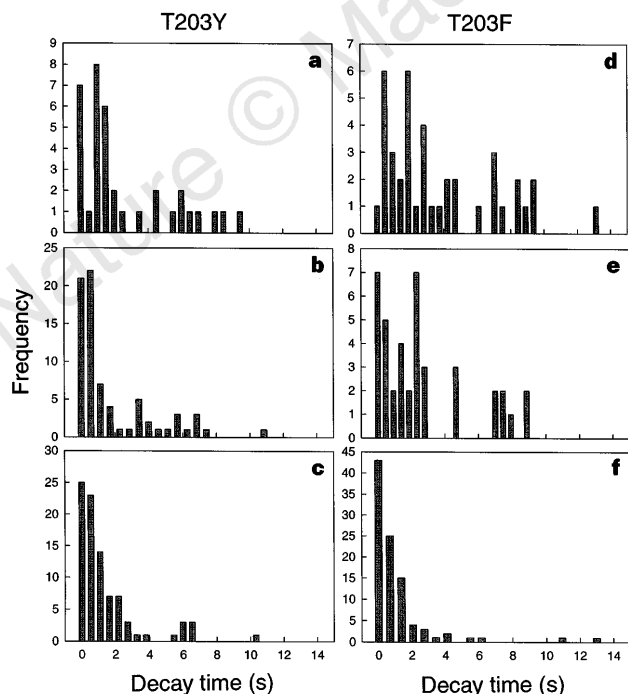


Figure 4 Histograms of single-molecule autocorrelation times τ_c as a function of incident intensity; average autocorrelation times, τ_c^{avg} , are calculated from the distributions. Histograms **a**, **b**, **c** correspond respectively to T203Y at 100 W cm^{-2} ($\tau_c^{\text{avg}} = 2.82 \pm 0.62 \text{ s}$, $n = 36$), 500 W cm^{-2} ($\tau_c^{\text{avg}} = 2.01 \pm 0.36 \text{ s}$, $n = 74$) and $2,000 \text{ W cm}^{-2}$ ($\tau_c^{\text{avg}} = 1.61 \pm 0.27 \text{ s}$, $n = 89$). Traces **d**, **e** and **f** were generated from T203F data at respectively 100 W cm^{-2} ($\tau_c^{\text{avg}} = 4.78 \pm 0.80 \text{ s}$, $n = 41$), 500 W cm^{-2} ($\tau_c^{\text{avg}} = 2.91 \pm 0.57 \text{ s}$, $n = 40$) and $2,000 \text{ W cm}^{-2}$ ($\tau_c^{\text{avg}} = 1.37 \pm 0.27 \text{ s}$, $n = 97$). Quoted errors correspond to 80% confidence limits.

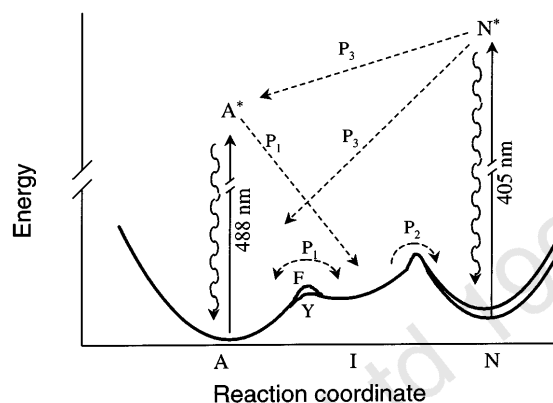


Figure 5 Schematic diagram of the various states consistent with the experimental observations as described in the text. For simplicity, the states are depicted as connected as $A \leftrightarrow I \leftrightarrow N$, but $I \leftrightarrow A \leftrightarrow N$ is also possible. A^* and N^* denote excited states of A and N, respectively. See text for definitions of other symbols.

The observed single-molecule dynamics and bulk absorption and emission spectra suggest the model in Fig. 5. Because T203Y and T203F have strong absorption maxima near 513 nm , we define the bright state as the anionic form, A, which absorbs 488-nm excitation and emits photons at 525 nm , and the dark states as those which do not. The blinking indicates that a dark state, I, exists from which spontaneous return to A occurs, although the exact identity of form I is unknown at present. Processes P_1 populate the state I, either directly, or as a result of thermally activated barrier-crossing produced by cooling of excited vibrational levels of the electronic ground state. The latter possibility is consistent with the weak power dependences observed in the τ_c values of both mutants (Fig. 4). In addition, there is another dark state which is eventually reached with lower probability (state N, process P_2) and which has a long lifetime ($>5 \text{ min}$) in the absence of light. We use “N” for this state as much evidence suggests this to be the neutral form^{15–18}, and the data in Fig. 1a suggests that the less-polar solvation shell produced with Phe 203 favours this neutral state. The ground-state barriers in both mutants for the transition from state N to state I are large enough to allow ambient stability of state N, consistent with the bi-exponential dark-time distributions. Excitation at 405 nm , however, does regenerate state A, by some third process P_3 , showing that light-induced conversion from the neutral to the anionic state occurs when higher-energy photons are present. Because the wild-type hydrogen-bonding and chromophore solvation shell are altered in T203Y and T203F, these mutants show distinct, but interconvertible, neutral and anionic states with very different photophysical properties.

These optical transformations of GFPs are likely to be obscured in ensemble-averaged measurements. Continuous monitoring of these immobilized single protein molecules shows that a process, which might be identified as irreversible ‘bleaching’, reverses on illumination at a different wavelength. This controllable photochromism may allow these GFP mutants to be used as caged fluorescent markers to monitor time-dependent cellular processes such as protein diffusion or transport²⁵. Photochromic GFPs could also offer potential advantages in optical information storage^{26,27} such as high storage densities (1 bit per molecule), photostability even in aereated samples, fluorescent readout, ease of production, known atomic resolution structure, and amenability to optimization by mutagenesis. □

Methods

Mutants. GFP was expressed in *Escherichia coli* using the expression plasmid pRSET (Invitrogen) in which the region encoding GFP was fused in-frame with

nucleotides encoding an amino-terminal polyhistidine tag. Sequence changes were introduced by site-directed mutagenesis using the Bio-Rad mutagenesis kit²⁸ and confirmed by sequencing. The recombinant proteins were expressed in the bacterial strain BL21(DE3) after induction with IPTG (0.5 mM) at room temperature and purified by Ni affinity chromatography.

T203F/T203Y imaging. Samples were prepared by the methods of ref. 21 from 10⁻¹⁰ M solutions of protein diluted in 1 mg ml⁻¹ BSA. Polyacrylamide gels ($T = 15\%$, $C = 5\%$ without SDS) were prepared in pH 7 phosphate-buffered saline doped with protein (here T is the total concentration of monomer in g per 100 ml, C is the wt% of total monomer which is crosslinker, and SDS indicates sodium dodecyl sulphate). The gel host provided pore sizes small enough for convenient (and complete) immobilization of each protein molecule, while maintaining its naturally fluorescent, native conformation^{21,29}. Excitation with a 488-nm laser (100–2,000 W cm⁻² at the gel/coverslip interface) occurred in the total-internal-reflection geometry; the emission was imaged with a Nikon inverted microscope (250-nm resolution) with an Omega 535DF55 filter and a Princeton Instruments intensified frame transfer CCD (100 ms time resolution, 90 s collection time). Oxygen was not removed from samples for which data is shown, but samples prepared with ~10 min helium bubbling showed similar on/off behaviour. 405-nm irradiation was produced by a Hg arc lamp with line filter through the epi-illumination port (1 W cm⁻²). The linear increase of detected photons as a function of laser intensity (100–2,000 W cm⁻²) indicated that saturation and multiphoton processes were negligible in these studies. Typical detected count rates of 5,000–6,000 photons s⁻¹ at 2,000 W cm⁻² pumping intensity (~150,000 excitations s⁻¹) were achieved, with most of the molecules emitting several millions of photons without irreversible bleaching.

Autocorrelation analysis. We define the autocorrelation function, $C(\tau)$, for discrete data points;

$$C(\tau) = \frac{\sum_{i=0}^N (I(i) - \bar{I})(I(i + \tau) - \bar{I})}{\sum_{i=0}^N (I(i) - \bar{I})^2}$$

where $\bar{I}$ is the average intensity, t is the time summed from 0 to N 100-ms intervals, and $I(t)$ is the time-dependent fluorescence intensity. Confidence limits were generated on the autocorrelations such that any values within the limits were consistent with zero³⁰. Exponential fits of the autocorrelations were generated only for the statistically significant portions of the curves beyond the short time correlation spike arising from band-limited noise.

Received 8 April; accepted 17 June 1997.

- Moerner, W. E. & Kador, L. Optical detection and spectroscopy of single molecules in a solid. *Phys. Rev. Lett.* **62**, 2535–2538 (1989).
- Orrit, M. & Bernard, J. Single pentacene molecules detected by fluorescence excitation in a p-terphenyl crystal. *Phys. Rev. Lett.* **65**, 2716–2719 (1990).
- Reilly, P. D. & Skinner, J. L. Spectral diffusion of single molecule fluorescence: a probe of low-frequency localized excitations in disordered crystals. *Phys. Rev. Lett.* **71**, 4257–4260 (1993).
- Moerner, W. E. Examining nanoenvironments in solids on the scale of a single, isolated molecule. *Science* **265**, 46–53 (1994).
- Barbara, P. F. & Moerner, W. E. (eds) Single molecules and atoms. *Acc. Chem. Res.* **29**, (1996).
- Betz, E. & Chichester, R. N. Single molecules observed by near-field scanning optical microscopy. *Science* **262**, 1422–1428 (1993).
- Ambrose, W. P., Goodwin, P. M., Martin, J. C. & Keller, R. A. Alterations of single molecule fluorescence lifetimes in near-field optical microscopy. *Science* **265**, 364–367 (1994).
- Nie, S., Chiu, D. T. & Zare, R. N. Probing individual molecules with confocal fluorescence microscopy. *Science* **266**, 1018–1021 (1994).
- Trautman, J. K. & Macklin, J. J. Time-resolved spectroscopy of single molecules using near-field and far-field optics. *Chem. Phys.* **205**, 221–229 (1996).
- Lu, H. P. & Xie, X. S. Single-molecule spectral fluctuations at room temperature. *Nature* **385**, 143–146 (1997).
- Funatsu, T., Harada, Y., Tokunaga, M., Saito, K. & Yanagida, T. Imaging of single fluorescent molecules and individual ATP turnovers by single myosin molecules in aqueous solution. *Nature* **374**, 555–559 (1995).
- Vale, R. D. *et al.* Direct observation of single kinesin molecules moving along microtubules. *Nature* **380**, 451–453 (1996).
- Schmidt, T., Schutz, G. J., Baumgartner, W., Gruber, H. J. & Schindler, H. Imaging of single molecule diffusion. *Proc. Natl Acad. Sci. USA* **93**, 2926–2929 (1996).
- Basché, T., Moerner, W. E., Orrit, M. & Wild, U. P. (eds) *Single Molecule Optical Detection, Imaging, and Spectroscopy* (Verlag-Chemie, Munich, 1997).
- Heim, R., Prasher, D. C. & Tsien, R. Y. Wavelength mutations and posttranslational autooxidation of green fluorescent protein. *Proc. Natl Acad. Sci. USA* **91**, 12501–12504 (1994).
- Cubitt, A. B. *et al.* Understanding, improving and using green fluorescent proteins. *Trends Biochem. Sci.* **20**, 448–455 (1995).
- Chatteraj, M., King, B. A., Blublitz, G. U. & Boxer, S. G. Ultra-fast excited state dynamics in green fluorescent protein: Multiple states and protein transfer. *Proc. Natl Acad. Sci. USA* **93**, 8362–8367 (1996).
- Brej, K. *et al.* Structural basis for dual excitation and photoisomerization of the *Aequorea victoria* green fluorescent protein. *Proc. Natl Acad. Sci. USA* **94**, 2306–2322 (1997).
- Ormo, M. *et al.* Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science* **273**, 1392–1395 (1996).

- Yang, F., Moss, L. G. & Phillips, J. G. N. The molecular structure of green fluorescent protein. *Nature Biotechnol.* **14**, 1246–1251 (1996).
- Dickson, R. M., Norris, D. J., Tzeng, Y.-L. & Moerner, W. E. Three dimensional imaging of single molecules in pores of poly(acrylamide) gels. *Science* **274**, 966–969 (1996).
- Moerner, W. E. *et al.* Optical probing of single molecules of terrylene in a Shpol'skii matrix—a two-state single-molecule switch. *J. Phys. Chem.* **98**, 7382–7389 (1994).
- Stuart, J. A., Tallent, J. R., Tan, E. H. L. & Birge, R. R. in *Sixth Biennial IEEE Intl Nonvolatile Memory Technol. Conf. Proc.* 35–51 (IEEE, Albuquerque, NM, 1996).
- Sakmann, B. & Neher, E. *Single Channel Recording* (Plenum, New York, 1995).
- Yokoe, H. & Meyer, T. Spatial dynamics of GFP-tagged proteins investigated by local fluorescence enhancement. *Nature Biotechnol.* **14**, 1252–1256 (1996).
- Irie, M. in *Photo-reactive Materials for Ultrahigh Density Optical Memory* (ed. Irie, M.) 1–12 (Elsevier Science, Amsterdam, 1994).
- Parthenopoulos, D. A. & Rentzepis, P. M. Three-dimensional optical storage memory. *Science* **245**, 843–845 (1989).
- Kunkel, T. A., Roberts, J. D. & Zakour, R. A. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Meth. Enzymol.* **154**, 367–382 (1987).
- Fawcett, J. S. & Morris, C. J. O. R. Molecular-sieve chromatography of proteins on granulated polyacrylamide gels. *Separat. Sci.* **1**, 9–26 (1966).
- Box, G. E. P. & Jenkins, G. M. *Time Series Analysis Forecasting and Control* (Holden-Day, San Francisco, 1970).

Acknowledgements. We thank L. S. B. Goldstein and S. Kummer for discussions. This work was supported by the NSF.

Correspondence and requests for materials should be addressed to W.E.M. (e-mail: wmoerner@ucsd.edu).

The soil production function and landscape equilibrium

Arjun M. Heimsath*, William E. Dietrich*, Kunihiko Nishiizumi† & Robert C. Finkel‡

* Department of Geology and Geophysics, 301 McCone Hall, University of California, Berkeley, California 94720-4767, USA

† Space Sciences Laboratory, University of California, Berkeley, California 94720-7450, USA

‡ Center for Accelerator Mass Spectrometry, Lawrence Livermore National Laboratory, Livermore, California 94550, USA

Hilly and mountainous landscapes are partially to completely covered with soil under a wide range of erosion and uplift rates, bedrock type and climate. For soil to persist it must be replenished at a rate equal to or greater than that of erosion. Although it has been assumed for over 100 years that bedrock disintegration into erodible soil declines with increasing soil mantle thickness^{1–9}, no field data have shown this relationship. Here we apply two independent field methods for determining soil production rates to hillslopes in northern California. First, we show that hillslope curvature (a surrogate for soil production⁷) varies inversely with soil depth. Second, we calculate an exponential decline of soil production rates with increasing soil depth from measurements of the *in situ* produced cosmogenic ¹⁰Be and ²⁶Al concentrations in bedrock sampled under soils of different depths. Results from both methods agree well and yield the first empirical soil production function. We also illustrate how our methods can determine whether a landscape is in morphological equilibrium or not.

Soil thickness depends on the balance between production and erosion of soil (Fig. 1). We define soil to be distinct colluvial material, lacking relict rock structure and derived from underlying bedrock. The mechanical disruption that destroys rock structure and lowers the soil–bedrock interface may be due to processes that are biotic (for example, burrowing and tree throw)⁷ or abiotic (for example, dissolution-induced collapse, freeze-thaw, and shear deformation). Downslope soil transport can occur by mass wasting, overland flow, and biogenic disturbance. We focus on diffusive transport where the sediment flux, $\bar{q}_s$, is proportional to slope, ∇z , such that $\bar{q}_s = -K\nabla z$; here K is equivalent to a diffusion coefficient with dimensions (length)² (time)⁻¹. This relationship was articulated by Davis¹⁰ and Gilbert¹¹ and has been extensively applied in

models of landscape evolution^{2,5,7,12–14}. It seems to be appropriate for biogenic transport^{7,15} and soil creep¹⁶.

If we use this relationship in the mass conservation equation for soil thickness, h (Fig. 1),

$$\rho_s \frac{\partial h}{\partial t} = -\rho_r \frac{\partial e}{\partial t} - \nabla \cdot \rho_s \tilde{q}_s \quad (1)$$

and assume steady-state soil thickness (that is, $\partial h/\partial t = 0$), then soil production $-(\partial e/\partial t)$ is given by

$$\frac{\partial e}{\partial t} = -\frac{\rho_s}{\rho_r} K \nabla^2 z \quad (2)$$

where ρ_s and ρ_r are soil and rock bulk densities, z is ground surface elevation, e is the elevation of the bedrock–soil interface and t is time. Under such conditions, soil production should depend on hillslope curvature, $\nabla^2 z$. Furthermore, if K is assumed to be spatially and temporally constant, we can use topographic curvature as a surrogate for soil production rates⁷. The form of the soil production function (that is, $-\partial e/\partial t = f(h)$) can then be defined with field measurements of curvature and depth. We note that, according to equation (2), spatial variation of curvature across diffusion-dominated regions of the landscape indicates spatial variation in local production rates and that the landscape is not in equilibrium.

We can test directly this depth-dependent soil production rate by adapting the cosmogenic nuclide method of determining erosion rates^{17,18}. If we assume that bedrock conversion to soil reaches a steady state under a constant soil thickness, h (and the soil bulk density remains constant), then the concentration of the cosmogenic radionuclide, C (in atoms g^{-1}), in the bedrock at the soil–bedrock interface is

$$C = P(h, \theta) \left(\frac{1}{\lambda + \frac{\rho_r \epsilon}{\Lambda}} \right) \quad (3)$$

where $P(h, \theta)$ is the nuclide production rate (in atoms $\text{g}^{-1} \text{yr}^{-1}$) at depth h and slope θ , Λ is the mean attenuation length ($\sim 165 \text{ g cm}^{-2}$), λ is the decay constant of the radionuclide ($\lambda = \ln 2/t_{1/2}$), and ϵ (in cm yr^{-1}) is $-\partial e/\partial t$ in equation (1). Equation (3) is of the same form as that used by others to calculate the erosion rate of bedrock (in which case, h and θ equal zero)^{17,19}. We solve for soil production rates as:

$$\epsilon = -\frac{\partial e}{\partial t} = \frac{\Lambda}{\rho_r} \left(\frac{P(h, \theta)}{C} \right) \quad (4)$$

The production rates of ^{10}Be and ^{26}Al in quartz are known^{20,21}, and we measure bulk densities and soil depths. The soil production function is determined by measuring the nuclide concentrations in bedrock sampled under different soil depths, calculating soil production rates, and plotting them against soil depth.

Equations (2) and (4) provide two independent methods to test the hypothesis that soil production rates decline with increasing soil thickness. The first relies entirely on field observations and can be applied only in areas where slope-dependent mass transport is dominant. The second does not require slope-dependent transport. Both require local soil thickness to be, on average, constant with time. The validity of this assumption will vary considerably depending on the field location.

We focused our study on small ridges ('noses') in Tennessee Valley, Marin County, California, a field site used for extensive geomorphological research^{7,22–25}. Intensely sheared thrust sheets of greenstone, greywacke sandstone and chert, typical of the Jurassic–Cretaceous Franciscan assemblage in the Marin Headlands terrane, underlie the field area²⁶. The area receives an average annual rainfall of 760 mm (ref. 27) and was grazed before 1972⁷. There is no evidence that Pleistocene climate variation caused

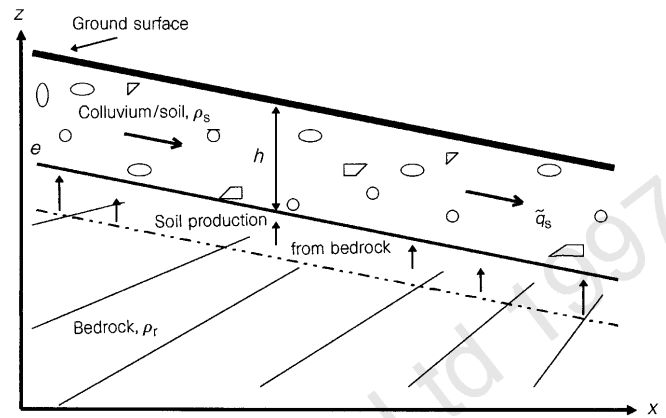


Figure 1 The conservation of mass equation for soil thickness h (equation (1)) states that the change in soil mass with time, t , is equal to the conversion of bedrock to soil due to lowering of the bedrock–soil interface less the divergence of transported soil mass. The area shown between the base of the soil at elevation e and the dashed line is the amount of bedrock that would be converted to soil over some specified time interval. In our study area, mass transport, $\tilde{q}_s$, of the entire active layer of soil is caused primarily by biogenic processes acting on an inclined surface. Here ρ_s and ρ_r are the densities of soil and bedrock, respectively. Note that $z = e + h$, a bedrock-fixed coordinate system not accounting for tectonic influences on absolute elevation, and h is much less than the landscape elevation scale. (Modified from Dietrich *et al.*⁷)

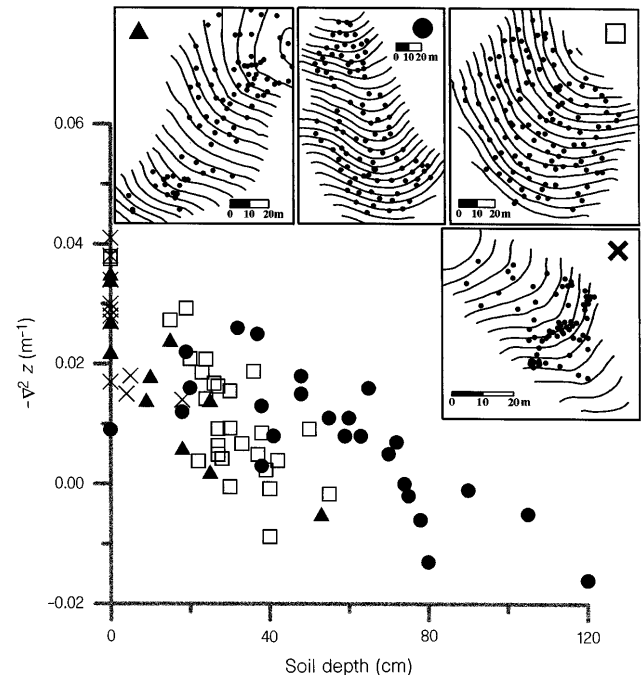


Figure 2 Plot of negative hillslope curvature, $-\nabla^2 z$, against local soil depth for four distinct small ridges (inset maps drawn with 2-m contour intervals) in Tennessee Valley, Marin County, California (37.9° N, 122.6° W). The large upper corner symbol on each inset map corresponds to the $-\nabla^2 z$ values plotted for the respective ridge and shows ridge locations on the landscape shown in Fig. 3 left inset. Dots on the inset maps are depth measurement locations. Each ridge was surveyed with 1–3 m resolution and gridded using a Krigging interpolation scheme. Curvature was calculated at each point using the interpolated elevations of the eight nearest grid neighbours. We found distinct soil–bedrock boundaries on each ridge. Soil depth was measured to the bottom of the colluvial layer, either by digging pits or by averaging several auger holes.

dramatically fluctuating hillslope erosion rates or processes, although net sediment storage in valleys and landslide frequency in unchannelled valleys (hollows) may have varied^{23,28}. Although there may have been partial Pleistocene forest cover, there is no evidence of a Holocene forest. The vegetation is a mixture of coastal grassland and scrub. In colluvium-mantled hollows, exfiltrating subsurface flow and rain on saturated areas generate extended areas of saturation overland flow²⁵. The soil mantle varies in thickness across the landscape and is typically an organic-rich, stony loam with weak to no horizon development. Soil production seems to be due primarily to biogenic disruption of weathered bedrock⁷. Soil and rock fragments from pocket gopher (*Thomomys bottae*) burrows abound, and soil pits show burrowing through the typically abrupt soil–bedrock boundary. Burrowing is also the primary mechanism for downslope soil transport²² and effective diffusivities have been quantified^{7,16,23}. Landsliding is mostly confined to steep hollows of thick colluvial deposits^{24,25}.

We selected four small ridges, each underlain by greywacke, for this study (Figs 2 and Fig. 3 left inset). The high topographic divergence of these small ridges and the modest slopes ensured that soil transport by landsliding has been insignificant. Figure 2 shows that curvature declines with increasing soil thickness, as predicted if curvature is a surrogate for soil production and soil production decreases with soil thickness. The large variance is expected because: (1) biogenic soil production causes short-term variation in local thickness; and (2) bedrock heterogeneity in resistance to weathering and mechanical disruption by biota leads to local variability in the curvature–thickness relationship. In general, strongly convex areas have thin soils and weakly convex areas have thicker soils.

We sampled bedrock from the base of the soil column and from exposed bedrock on three of the four surveyed small ridges and at other locations for cosmogenic nuclide analyses (Fig. 3 left inset). The samples reflect the range of soil thickness measured on these ridges. Figure 3 main figure shows that soil production rates, calculated from both ¹⁰Be and ²⁶Al concentrations, decline exponentially with increasing soil depth. Erosion rates of prominent, isolated bedrock outcrops varied with rock type, with chert eroding the slowest. The average erosion rates we measured from stream sediments^{29,30} for two steep tributaries (labelled 1 and 2 in Fig. 3 left inset) were close to the maximum soil production rates (Table 1). Figure 3 right inset shows good agreement between the two methods even though no parameters were adjusted to match these data.

Two general production laws have been proposed³¹. An exponential decline of soil production with increasing soil depth was assumed to simulate the decrease in effectiveness of such mechan-

ical processes as freeze-thaw³ or biogenic disturbance⁷. A complex, bell-shaped polynomial function follows the intuition that maximum soil production occurs under a thin layer of soil^{3,8}. Dietrich *et al.*⁷ point out, however, that soil depths less than the peak are unstable (perturbations in thickness would strip the soil to bedrock) and found field-based agreement with an exponential production function. Although our data cannot reject the polynomial function, they do show that the peak production rate would occur under near-zero soil depth and would make little difference to modelling⁹.

An important assumption in our analysis is steady-state local soil depth during soil production and transport. This assumption is justified in several ways. There is no evidence for shallow landsliding or erosion by overland flow on the convex regions we studied. Soil production by burrowing tends not to alter local soil thickness dramatically during burrowing. Numerical experiments by Dietrich *et al.*⁷ show that diffusion-dominated ridges quickly (in a few thousand years) reach local steady-state soil thickness. Topographic change on such ridges is slow and changes in soil production or transport with Pleistocene–Holocene climate change would have adjusted local soil thickness early in the Holocene. Last, the observed ¹⁰Be and ²⁶Al concentrations imply erosion rates that would result in steady-state radionuclide concentrations in a few tens of thousands of years.

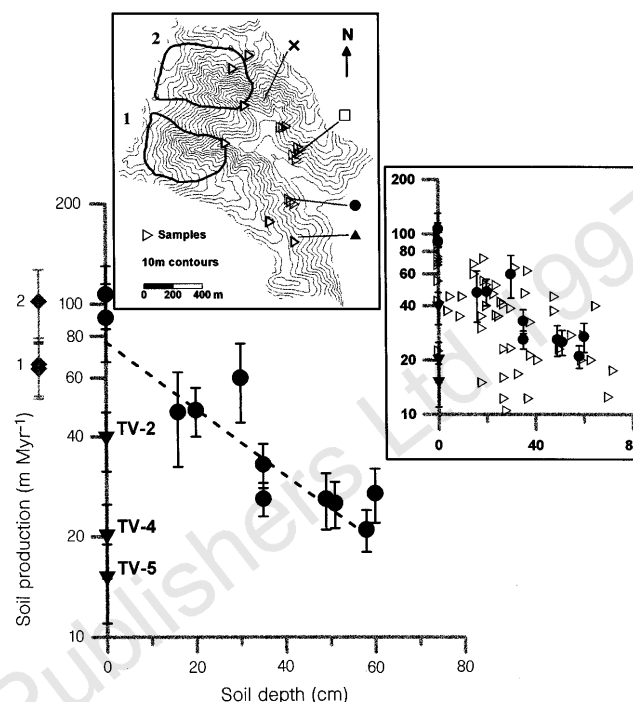
In contrast to the local steady-state condition for soils on the small ridges, our observations indicate that the landscape may be slowly changing morphologically. We find both the thinnest soils and the highest production rates on ridge crests, indicating the greatest lowering rates. Individual ridges with their different degrees of curvature may also be lowering at different rates (Fig. 2). Erosion rate varies with lithology and solitary outcrops are eroding slower than the surrounding landscape. Basin-wide erosion rates in the two primary tributaries of our catchment suggest that erosion is higher in these steeper regions than on the gentler, lower-elevation small ridges; Dietrich *et al.*⁷ observed that nearly all shallow landslide scars in the catchment are in those regions. The northern tributary basin may be lowering more rapidly than the southern one (Fig. 3). Although local, short-term variation in lowering rates can be expected on landscapes, our results suggest that Tennessee Valley is undergoing systematic morphological change. We also observe that such deviation from dynamic equilibrium can be inferred from landscape form, as proposed by Dietrich *et al.*⁷ For the divergent ridge crests, or ridges where diffusive transport processes are dominant, spatial variation in curvature implies variation in lowering rates. We note however, that soil depth variation alone does not indicate disequilibrium. Ahnert⁹ showed that constant-form

Table 1 Measurements of cosmogenic nuclide concentrations

Sample	Depth (cm)	Slope (deg)	Elev. (m)	Quartz wt (g)	²⁶ Al (10 ⁶ atoms g ⁻¹)	¹⁰ Be (10 ⁶ atoms g ⁻¹)	²⁶ Al/ ¹⁰ Be	<i>h</i> -slope factor	$-ae/at$ (m Myr ⁻¹)
TV-2	0	8	135	40.65	0.600 ± 0.101	0.115 ± 0.006	5.22 ± 0.92	1	39 ± 8
TV-3	16	10	120	40.06	0.447 ± 0.054	0.108 ± 0.007	4.15 ± 0.56	0.87	47 ± 15
TV-4	0	15	275	42.68	1.132 ± 0.078	0.248 ± 0.010	4.55 ± 0.37	0.98	20 ± 5
TV-5	0	0	275	40.22	1.433 ± 0.121	0.351 ± 0.017	4.08 ± 0.40	1.00	15 ± 4
TV-6	35	15	105	42.03	1.104 ± 0.050	0.195 ± 0.008	5.65 ± 0.34	0.69	26 ± 3
TV-7	58	20	100	42.58	1.446 ± 0.061	0.229 ± 0.016	6.32 ± 0.52	0.52	21 ± 3
TV-10	51	17	115	40.71	0.939 ± 0.104	0.171 ± 0.011	5.49 ± 0.71	0.59	25 ± 4
TV-11	0	21	120	32.39	0.234 ± 0.035	0.040 ± 0.005	5.92 ± 1.20	0.98	107 ± 23
TV-12	30	15	116	40.56	0.394 ± 0.091	0.074 ± 0.006	5.28 ± 1.30	0.72	60 ± 16
TV-13	49	18	140	40.03	1.060 ± 0.112	0.151 ± 0.013	7.03 ± 0.95	0.61	26 ± 5
TV-15	20	15	135	40.33	0.563 ± 0.055	0.083 ± 0.006	6.77 ± 0.83	0.85	48 ± 8
TV-16	35	20	133	45.05	0.741 ± 0.076	0.134 ± 0.007	5.54 ± 0.64	0.68	33 ± 5
TV-17	60	25	133	52.57	0.914 ± 0.093	0.161 ± 0.015	5.68 ± 0.77	0.54	27 ± 5
TV-23	0	15	137	40.28	0.260 ± 0.051	0.050 ± 0.005	5.16 ± 1.14	0.98	91 ± 24
Creek1a	NA	NA	110	51.50	0.414 ± 0.053	0.063 ± 0.005	6.61 ± 1.00	1	64 ± 12
Creek1b	NA	NA	110	40.22	0.366 ± 0.050	0.063 ± 0.006	5.81 ± 0.95	1	66 ± 13
Creek2	NA	NA	110	58.22	0.239 ± 0.049	0.041 ± 0.005	5.87 ± 1.37	1	102 ± 25

Concentration errors include 1σ from accelerator mass spectrometry. All errors are propagated to $-ae/at$, where *e* is the elevation of the bedrock–soil interface and *t* is time. Average soil density, 1.4 g/cm³; location is at 37.9°N, 122.6°W. ²⁶Al and ¹⁰Be production rates are corrected for elevation and location^{20,21}. The *h*-slope factor corrects for soil depth and slope (D. Lal, personal communication). NA, not available.

Figure 3 Main figure shows soil production rates ($-\partial e/\partial t$) (calculated from *in situ* produced cosmogenic ^{10}Be and ^{26}Al in bedrock samples) versus measured soil depths h . Al and Be were chemically separated and purified²² after chemically separating quartz and adding Be carrier; we measured ^{10}Be and ^{26}Al concentrations at the Lawrence Livermore National Laboratory accelerator mass spectrometry (AMS) facility²³. We normalized the measured ratios to the ICN ^{10}Be and the NBS ^{26}Al standards (Table 1). We plot the average calculated rate from both nuclides. Left inset, map showing cosmogenic nuclide sample locations (open triangles) in Tennessee Valley, Marin County, California. Samples (plotted in the main figure as filled circles) were taken from bedrock exposed at the ground surface or the soil-bedrock boundary revealed at the base of pits dug across three of the four ridges shown in Fig. 2. The exponential fit to these data is $-\partial e/\partial t = (77 \pm 9)e^{(-0.023 \pm 0.003)h}$. Other outcrops (plotted in the main figure as upside-down triangles) were sampled from a large, ~25-m relief, greenstone (TV-4), a 3-m-high bedded chert (TV-5) at the top of sub-basin 2, and from outcropping greywacke (TV-2) at one of the surveyed hillslope ridge crests. Two different creek sediment samples (1 and 2, filled diamonds, plotted to left of y axis in the main figure) were taken to estimate average erosion rates^{29,30}. Right inset, plot showing an overlay of the nuclide-based results with the soil production rates calculated from ^{-2}Z values (Fig. 2) using a regional diffusivity of $50 \text{ cm}^2 \text{ yr}^{-1}$ (refs 7, 16, 23) and average $\rho_s/\rho_r = 0.5$. In three cases, TV-3, TV-4 and TV-5, the observed $^{26}\text{Al}/^{10}\text{Be}$ ratio differs significantly from the expected production ratio of 6.0. These disparities may indicate either that the samples have experienced complex exposure histories or that some analytical problem exists. Here we interpret these samples with our simple exposure model, expressing the disparity between the ^{26}Al and the ^{10}Be results with the large uncertainty in the soil production/erosion estimates for these samples. This uncertainty does not alter the inference that the bedrock outcrops are eroding more slowly than other parts of the landscape. Only TV-3 is used in determining the soil production curve and, in fact, deleting it would not change the function.



hillslopes with spatially variable soil depths develop when the soil production function varies with underlying lithology.

The peak soil-production rate determines under what tectonic and erosional environments bedrock will emerge or become dominant across a landscape. If the soil production function varies with climate, bedrock and biota, then landscapes under similar tectonic regimes can have different morphology, ranging from fully soil-mantled rounded hills to bedrock cliffs. Our two independent methods provide a means to define empirically soil production functions for different rock types and climates. This should facilitate exploring the role of geology and climate in landscape evolution. □

Received 28 January; accepted 12 June 1997.

- Gilbert, G. K. *Report on the Geology of the Henry Mountains* (US Geol. Surv., Washington DC, 1877).
- Culling, W. E. H. Soil creep and the development of hillslope slopes. *J. Geol.* **71**, 127–161 (1963).
- Ahnert, F. In *The Evolution des Versants* (ed. Macar, P.) 23–41 (Univ. Liege, 1967).
- Tucker, G. E. & Slingerland, R. Predicting sediment flux from fold and thrust belts. *Basin Res.* **8**, 329–349 (1996).
- Anderson, R. S. & Humphrey, N. F. In *Quantitative Dynamic Stratigraphy* (ed. Cross, T.) 349–361 (Prentice Hall, Englewood Cliffs, New Jersey, 1989).
- Rosenbloom, N. A. & Anderson, R. S. Hillslope and channel evolution in a marine terraced landscape, Santa Cruz, California. *J. Geophys. Res.* **99**, 14013–14029 (1994).
- Dietrich, W. E., Reiss, R., Hsu, M.-L. & Montgomery, D. R. A process-based model for colluvial soil depth and shallow landsliding using digital elevation data. *Hydrol. Proc.* **9**, 383–400 (1995).
- Carson, M. A. & Kirkby, M. J. *Hillslope Form and Process* (Cambridge Univ. Press, New York, 1972).
- Ahnert, F. Approaches to dynamic equilibrium in theoretical simulations of slope development. *Earth Surf. Proc. Landforms* **12**, 3–15 (1987).
- Davis, W. M. The convex profile of badland divides. *Sciences* **20**, 245 (1892).
- Gilbert, G. K. The convexity of hilltops. *J. Geol.* **17**, 344–350 (1909).
- Kirkby, M. J. *Hillslope Process-response Models Based on the Continuity Equation* 15–30 (Spec. Publ. 3, Inst. Brit. Geogr., London, 1971).
- Koons, P. O. The topographic evolution of collisional mountain belts; a numerical look at the Southern Alps, New Zealand. *Am. J. Sci.* **289**, 1041–1069 (1989).
- Howard, A. D. A detachment-limited model of drainage basin evolution. *Wat. Resour. Res.* **30**, 2261–2285 (1994).
- Reneau, S. L. & Dietrich, W. E. Erosion rates in the southern Oregon Coast Range: evidence for an equilibrium between hillslope erosion and sediment yield. *Earth Surf. Proc. Landforms* **16**, 307–322 (1991).
- McKean, J. A., Dietrich, W. E., Finkel, R. C., Southon, J. R. & Caffee, M. W. Quantification of soil

- production and downslope creep rates from cosmogenic ^{10}Be accumulations on a hillslope profile. *Geology* **21**, 343–346 (1993).
- Lal, D. & Arnold, J. R. Tracing quartz through the environment. *Proc. Indian Acad. Sci. (Earth Planet. Sci.)* **94**, 1–5 (1985).
- Nishiizumi, K., Lal, D., Klein, J., Middleton, R. & Arnold, J. R. Production of ^{10}Be and ^{26}Al by cosmic rays in terrestrial quartz *in situ* and implications for erosion rates. *Nature* **319**, 134–136 (1986).
- Nishiizumi, K. *et al.* Cosmic ray produced ^{10}Be and ^{26}Al in Antarctic rocks: exposure and erosion history. *Earth Planet. Sci. Lett.* **104**, 440–454 (1991).
- Lal, D. Cosmic ray labeling of erosion surfaces: *in situ* nuclide production rates and erosion models. *Earth Planet. Sci. Lett.* **104**, 424–439 (1991).
- Nishiizumi, K. *et al.* Cosmic ray production rates of ^{10}Be and ^{26}Al in quartz from glacially polished rocks. *J. Geophys. Res.* **94**, 17907–17915 (1989).
- Black, T. A. & Montgomery, D. R. Sediment transport by burrowing animals, Marin County, California. *Earth Surf. Proc. Landforms* **16**, 163–172 (1991).
- Reneau, S. L. Depositional and erosional history of hollows; application to landslide location and frequency, long-term erosion rates, and the effects of climatic change. Thesis, Univ. California, Berkeley (1988).
- Dietrich, W. E., Wilson, C. J., Montgomery, D. R., McKean, J. & Bauer, R. Erosion thresholds and land surface morphology. *Geology* **20**, 675–679 (1992).
- Dietrich, W. E., Wilson, C. J., Montgomery, D. R. & McKean, J. Analysis of erosion thresholds, channel networks, and landscape morphology using a digital terrain model. *J. Geol.* **101**, 259–278 (1993).
- Wahrhaftig, C. Structure of the Marin headlands block, California: a progress report. *Franciscan Geol. N. California* **43**, 31–50 (1984).
- Rantz, S. E. Average annual precipitation and runoff in north coastal California. (Hydrol. Atlas 298, scale 1:1,000,000, US Geol. Surv. Menlo Park, 1968).
- Rypins, S., Reneau, S. L., Byrne, R. & Montgomery, D. R. Palynologic and geomorphic evidence for environmental change during the Pleistocene-Holocene transition at Point Reyes Peninsula, central coastal California. *Quat. Res.* **32**, 72–87 (1989).
- Granger, D. E., Kirchner, J. W. & Finkel, R. Spatially averaged long-term erosion rates measured from *in situ*-produced cosmogenic nuclides in alluvial sediment. *J. Geol.* **104**, 249–257 (1996).
- Bierman, P. & Steig, E. J. Estimating rates of denudation using cosmogenic isotope abundances in sediment. *Earth Surf. Proc. Landforms* **21**, 125–139 (1996).
- Cox, N. J. On the relationship between bedrock lowering and regolith thickness. *Earth Surf. Proc. Landforms* **5**, 271–274 (1980).
- Kohl, C. P. & Nishiizumi, K. Chemical isolation of quartz for measurement of *in situ*-produced cosmogenic nuclides. *Geochim. Cosmochim. Acta* **56**, 3585–3587 (1992).
- Davis, J. C. *et al.* LLNL/UC AMS facility and research program. *Nucl. Instrum. Meth. B* **52**, 269–272 (1990).

Acknowledgements. We thank K. Heimsath and L. Cossey for field and laboratory assistance; D. DePaolo for laboratory space; D. Lal for suggestions; and the Golden Gate National Recreation Area for access to our study site. We thank P. Bierman and A. Howard for comments on the manuscript. This work was supported by Cal Space, IGPP-LLNL, NSF, NASA, DOE and a Switzer Environmental fellowship.

Correspondence and requests for materials should be addressed to A.M.H. (e-mail: arjun@moray.berkeley.edu).

Experimental and theoretical identification of a new high-pressure phase of silica

L. S. Dubrovinsky*, S. K. Saxena*, P. Lazor*, R. Ahuja†, O. Eriksson†, J. M. Wills‡ & B. Johansson†

* Theoretical Geochemistry Program, Institute of Earth Sciences, Uppsala University, S-752 36, Uppsala, Sweden

† Condensed Matter Theory Group, Department of Physics, Uppsala University, Box 530, S-751 21, Uppsala, Sweden

‡ Theoretical Division, Los Alamos Laboratory, Los Alamos, New Mexico 87545, USA

Following the discovery of stishovite (the highest-pressure polymorph of silica known from natural samples), many attempts have been made to investigate the possible existence of denser phases of silica at higher pressures. Based on the crystal structures observed in chemical analogues of silica^{1–3}, high-pressure experiments on silica^{4–11} and theoretical studies^{12–16}, several possible post-stishovite phases have been suggested. But the likely stable phase of silica at pressures and temperatures representative of Earth's lower mantle remains uncertain. Here we report the results of an X-ray diffraction study of silica that has been heated to temperatures above ~2,000 K and maintained at pressures between 68 and 85 GPa. We observe the occurrence of a new high-pressure phase which we identify with the aid of first-principles total-energy calculations. The structure of this phase (space group *Pnc2*) is intermediate between the α -PbO₂ and ZrO₂ structures, and is denser than other known silica phases.

Experiments at extreme conditions are usually associated with significant difficulties; these include limited size of sample (and bad access to it) in a diamond-anvil cell, non-quenchable phase transitions, and pressure and temperature gradients. The available experimental data on SiO₂ at mantle pressures are either from unheated samples^{6,9–11} or from heated samples quenched at high pressures⁸, and then there are only few lines observed in the X-ray diffraction spectra. The present experimental study is designed to minimize the difficulties encountered in earlier studies. We used two different starting materials (silica gel and quartz), laser-heated these materials at various pressures and followed the transformations both with increasing and decreasing pressure. We used the Mao–Bell type of diamond-anvil cell following the sample preparation method described elsewhere¹⁷. We conducted three separate sets of experiments.

In the first set of experiments, a 10- μ m-thick iron foil was surrounded by dried silica gel. In the two other sets of experiments, we used quartz and platinum in the same manner. In the first set of experiments, an area of 60 μ m diameter was heated to 2,000($\pm$ 50) K at ~80 GPa with a Nd:YAG laser for several dozen minutes. In the experiments with quartz and platinum, samples were heated over an area of 80 μ m diameter for at least 30–40 min at different pressures (40–68 GPa). The temperature was measured using the spectroradiometric method¹⁷. The quenched samples, still under pressure, were studied at the Brookhaven National Laboratory synchrotron X-ray facility or the X-ray facilities in Uppsala University. The pressures were determined using the equation of state^{18,19}. There is a close agreement ($\pm$ 2 GPa) between the pressures obtained from different pressure scales. The pressure was rather homogenous within the studied areas (maximum difference; 5 GPa at a pressure of 80 GPa). An estimated thermal pressure value of ~7% (ref. 17) has been added to the reported pressures.

After heating at a pressure of 85($\pm$ 5) GPa in the experiment with iron and silica gel and at 68($\pm$ 5) GPa in the experiments with quartz and platinum, the silica closest to the metal was found to have

Table 1 Assignments of reflections and comparison between observed and calculated *d* values

d_{obs} (Å)	I (%)	Metal		CaCl ₂ -like phase		New silica phase	
		hkl	d_{calc}	hkl	d_{calc}	hkl	d_{calc}
85(±5) GPa, iron and silica gel are starting materials							
2.9905	11					011	2.9980
2.8014	8			110	2.8051		
2.4615	14					111	2.4590
2.1024	10			011	2.0999		
2.0090	35	100	2.0093				
1.882*	20			111	1.8713	210	1.8827
1.865*	20	002	1.8500				
1.820*	25			210	1.8155	112	1.8203
1.7666	100	101	1.7662				
1.444*	10					220	1.4442
1.435*	10			121	1.4278		
1.418*	17					122	1.4154
1.407*	20			220	1.4026		
1.2928	12			310	1.2940	311	1.2931
1.2232	11			221	1.2246	302	1.2229
68(±5) GPa, platinum and quartz are starting materials							
3.0130	19					0.11	3.0133
2.7881	91			110	2.7873		
2.4719	52					111	2.4714
2.1806	16			101	2.1803		
2.1446	53	111	2.1449				
2.0685	6					102	2.0681
1.8942	47			111	1.8969	210	1.8918
1.8578	100	200	1.8575				
1.8296	16					112	1.8291
1.742*	11			120	1.7373	211	1.7555
1.4716	22			211	1.4721		
1.457*	40			121	1.4426	013	1.4577
1.4220	33					122	1.4226
1.3940	13			220	1.3940		
1.3132	24	220	1.3135				
1.2986	17			002	1.2945	311	1.2993

Iron at 85($\pm$ 5) GPa is indexed as the ϵ -phase. The new silica phase is indexed as orthorhombic with lattice parameters $a = 4.300(\pm 1)$ Å; $b = 3.899(\pm 2)$ Å; $c = 4.689(\pm 4)$ Å, $V = 11.84(\pm 2)$ cm³ mol⁻¹ at 68($\pm$ 5) GPa and $a = 4.320(\pm 2)$ Å; $b = 3.919(\pm 2)$ Å; $c = 4.711(\pm 3)$ Å, $V = 12.01(\pm 2)$ cm³ mol⁻¹ at 85($\pm$ 5) GPa. Comparison between the molar volume of the CaCl₂-like phase ($a = 4.043(\pm 3)$ Å; $b = 3.848(\pm 2)$ Å; $c = 2.589(\pm 2)$ Å, $V = 12.13(\pm 2)$ cm³ mol⁻¹ at 68($\pm$ 5) GPa) and the molar volume of the new silica phase ($a = 4.125(\pm 8)$ Å; $b = 3.826(\pm 7)$ Å; $c = 2.512(\pm 8)$ Å, $V = 11.94(\pm 5)$ cm³ mol⁻¹ shows that the latter is ~1% more dense. *I* indicates relative intensity.

Table 2 Theoretically optimized structural parameters of the *Pnc2* phase of silica

Volume (Å ³ per formula unit)	<i>b/a</i>	<i>c/a</i>	Atomic coordinates			
			Si1	Si2	O1	O2
19.57	0.898	1.114	0.0 0.143	0.5 0.832	0.833 0.260 0.377	0.667 0.757 0.104
16.31	0.899	1.115	0.0 0.140	0.5 0.836	0.832 0.263 0.378	0.668 0.759 0.103

The FPLMTO method (see Fig. 3 legend) was used to optimize the parameters for two different volumes.

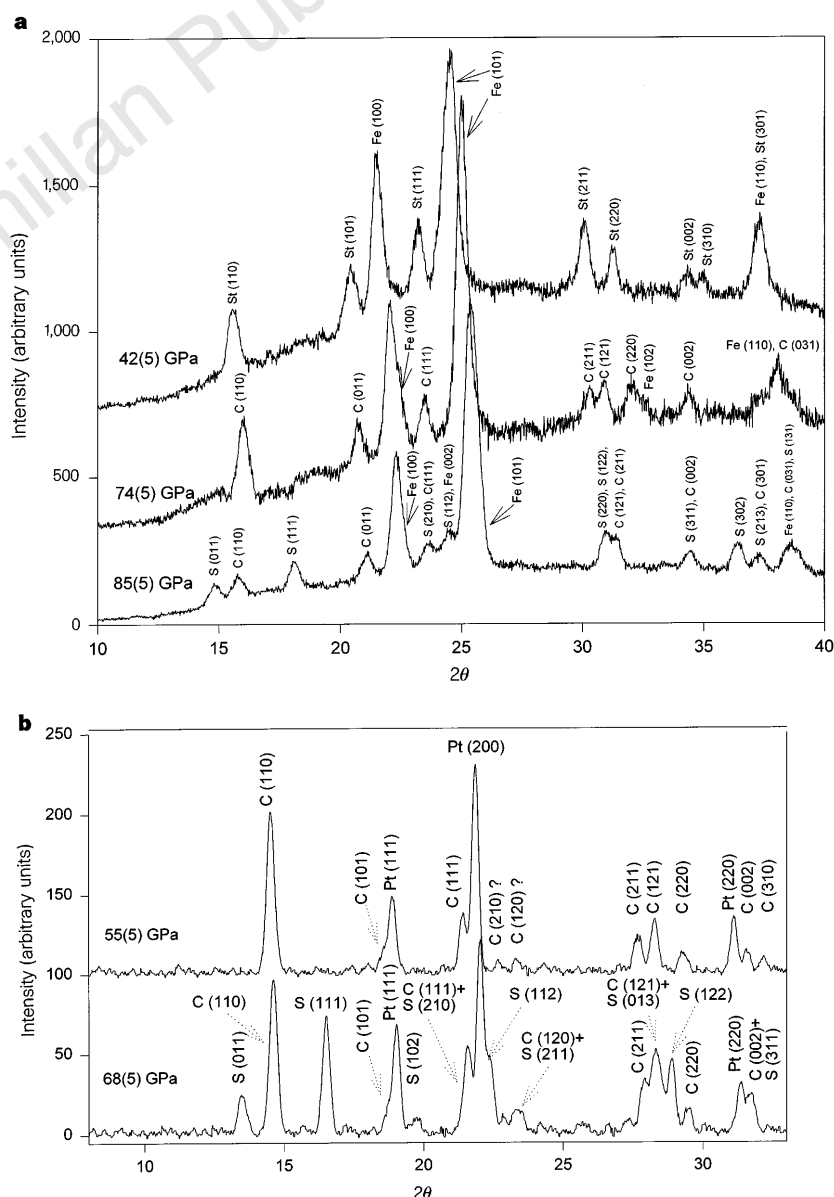
recrystallized to a dense form of SiO_2 over large areas (Figs 1, 2 and Table 1). X-ray diffraction shows a mixture of three phases: metal (ϵ -Fe or Pt), a CaCl_2 -like phase and the new silica phase. The metal reflections are quite strong in all patterns and can be easily identified (Figs 1, 2). Some reflections can be explained by the presence of CaCl_2 -like structure in the sample (Table 1). This happens because the heating produced by the Nd YAG laser is very localized; only the part of SiO_2 which is in contact with the metal foil (acting as a hot-plate) has the same temperature as the foil. Away from the heated area both in the axial and radial directions, the temperature decreases rapidly. Therefore, while stishovite (or the CaCl_2 -like phase) in the immediate surroundings of the metal would be converted to a new phase, the temperature away from the metal foil will be lower; in these latter areas stishovite (or the CaCl_2 -like phase) will remain stable.

A number of diffraction lines (for example, lines with d values of 2.9905, 2.4615 and 1.4181 Å at $85(\pm 5)$ GPa in the experiment with silica gel or lines at 3.0130, 2.4719, 2.0685, 1.8296 and 1.4220 Å at $68(\pm 5)$ GPa in the experiment with quartz) belong neither to the metal nor to the CaCl_2 -like phase (or stishovite) and must be considered as being due to the new phase (Table 1). These lines cannot belong to any iron phase, because they are absent in the

diffraction patterns of those parts of the sample that consist mainly of ϵ -Fe, and they appear only in the presence of silica. Moreover, they appear also in experiments with platinum. We cannot explain these reflections as being due to either pressure gradients or some reaction between metal and SiO_2 . After collecting the data at high pressure, we released the pressure gradually and noted the disappearance of the peaks of the new silica phase (Fig. 1).

Different investigators have suggested a number of possible post-stishovite phases of SiO_2 . Such phases have been obtained from theoretical simulations or observed experimentally in related systems; these are CaCl_2 , α - PbO_2 and modified α - PbO_2 (space group $I2/a$), ZrO_2 , intermediate between α - PbO_2 and ZrO_2 (space group $Pnc2$), α - PbCl_2 , fluorite and modified fluorite (space group $Pa3$) structures^{1–16}. We have considered the diffraction patterns of these phases and tried to index our experimental diffraction pattern accordingly. It was found that all reflections in our diffraction pattern can be easily explained by the presence of a mixture of metal, a CaCl_2 -like phase and a structure with $Pnc2$ symmetry. At high pressure α - PbO_2 and $Pnc2$ structures are similar^{15,16} and it is difficult to distinguish these two structures on the basis of our experimental data. But on the basis of theoretical considerations (see below) the $Pnc2$ structure is preferable.

Figure 1 a, X-ray diffraction pattern of the sample containing silica gel and iron as starting materials after laser heating during decompression. The sample was heated at $2,000(\pm 100)$ K. The initial diffraction pattern at $85(\pm 5)$ GPa contains lines of a new phase with $Pnc2$ structure (marked as S), CaCl_2 -like silica phase (C) and ϵ -iron (Fe). At $74(\pm 5)$ GPa the reflections of the $Pnc2$ structure almost completely disappear owing to conversion to the CaCl_2 -like silica phase with lattice parameters $a = 4.041(\pm 3)$ Å, $b = 3.846(\pm 4)$ Å, $c = 2.588(\pm 2)$ Å. At lower pressure, $42(\pm 5)$ GPa, the CaCl_2 -like silica phase transforms to stishovite ($a = 4.029(\pm 4)$ Å, $c = 2.595(\pm 6)$ Å). The pressures were determined on the bases of the iron equation of state^{17,18}. **b**, X-ray diffraction pattern of the sample, containing silica gel and iron as starting materials, after laser heating during decompression. The bottom curve is from the sample heated to $2,450(\pm 150)$ K during 40 min at $68(\pm 5)$ GPa. The diffraction pattern shows lines of the new phase $Pnc2$ (marked as S), CaCl_2 -like silica phase (C) and platinum (Pt). The upper curve corresponds to the case when the pressure was decreased to $55(\pm 5)$ GPa and platinum heated again to about $1,500(\pm 200)$ K for nearly 40 min by scanning with the laser. Then only the CaCl_2 -like ($a = 4.071(\pm 4)$ Å, $b = 3.876(\pm 4)$ Å, $c = 2.607(\pm 3)$ Å) reflections remain. These experiments establish the reversibility of the transformation.



We have also studied the stishovite (or CaCl_2 -type)– $Pnc2$ structural transition by means of accurate first-principles, total-energy calculations. This is necessary because our interpretation of the diffraction patterns is based on the results from lattice dynamics simulations using semi-empirical inter-atomic potentials^{14,15}. Only *ab initio* calculations can confirm results of such semi-empirical

methods. Figure 3a shows the results of our present first-principles theoretical total-energy calculations; Fig. 3b shows the difference in total energies between different structures as a function of volume and where the $Pnc2$ structure is used as a reference energy level. One can see from Fig. 3b that ideal α - PbO_2 is clearly not stable in any region over the whole studied volume range. In the lattice dynamics

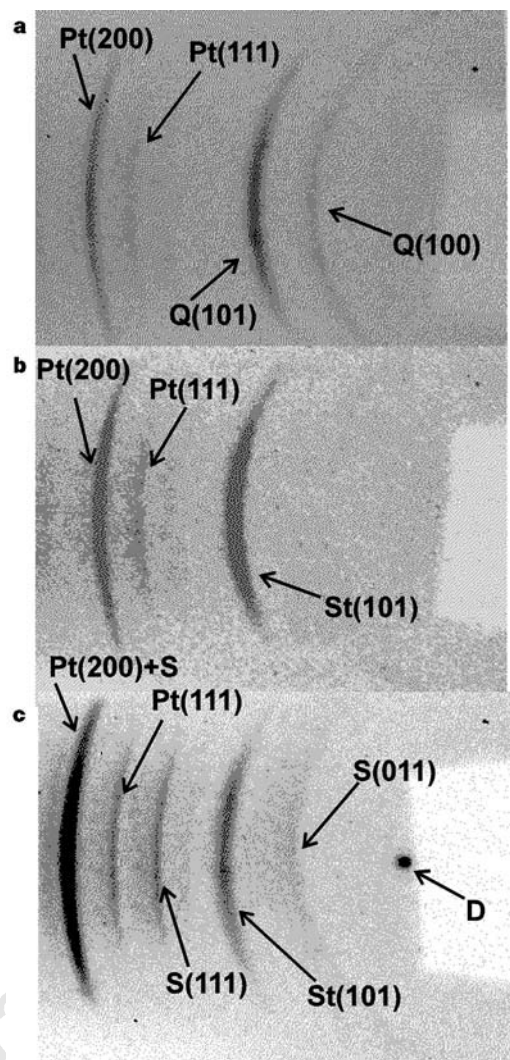


Figure 2 Examples of diffraction patterns collected during the transformation of quartz at high pressure after heating, using a CCD (charge-coupled device) area detector and monochromatic $\text{Mo K}_{\alpha 1,2}$ radiation (rotation anode generator, 18 kW). The irradiated area corresponds to $50 \mu\text{m}$ and the exposure time is 6 hours. **a**, Quartz at $45(\pm 5)$ GPa before heating. The letter Q denotes quartz (or quartz II^{H}) reflections. Owing to preferred orientation, the (200) reflection of Pt is more intense than the (111) reflection. **b**, Same sample after heating at $2,350(\pm 150)$ K using a Nd:YAG laser. The quartz reflections have disappeared and the (101) reflection of stishovite (St) has appeared. Note that we cannot distinguish between the stishovite and CaCl_2 structures within the angle range shown here. **c**, Same sample after pressurizing at $68(\pm 5)$ GPa and heating at $2,450(\pm 150)$ K. New reflections at $3.02 \text{ \AA}$ and $2.48 \text{ \AA}$ appear. These reflections can be indexed as (011) and (111) of the $Pnc2$ structure (S). Letter D denotes a single spot from the diamond anvil.

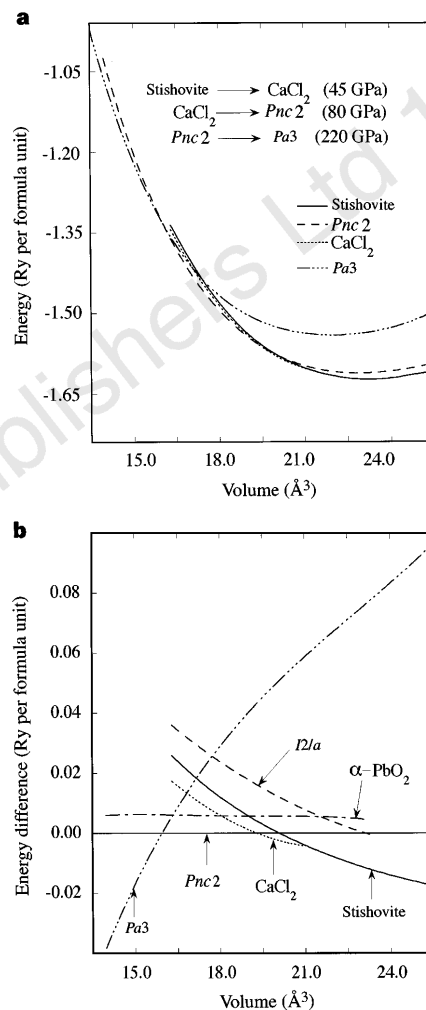


Figure 3 a, Calculated total energies (relative to $E = -876.000 \text{ Ry}$) of the stishovite, CaCl_2 , $Pnc2$ and $Pa3$ crystal structures for SiO_2 as a function of volume. From the standard 'common tangent' construction we derive the following volume collapses at the structural transitions; $\Delta V/V = 1.5\%$ for $\text{CaCl}_2 \rightarrow Pnc2$ and $\Delta V/V = 3.5\%$ for $Pnc2 \rightarrow Pa3$. The theoretical technique used here is a so-called full potential linear muffin-tin orbital method (FPLMTO)²¹. The only approximation which enters this theory is the use of the local density approximation to the exchange-correlation potential (Hedin-Lundqvist parametrization²²). The truncation of the expansion of wavefunctions, charge density and potential is also an approximation but the numerical error introduced by this truncation can always be reduced to a desired small value. Thus a high level of numerical accuracy has been maintained throughout the calculations. The basis functions, electron densities, and potentials were expanded in combinations of spherical harmonic functions (with a cut-off $l_{\text{max}} = 8$) inside non-overlapping spheres surrounding the atomic sites (muffin-tin spheres) and in a Fourier series in the interstitial region. The muffin-tin sphere occupies $\sim 40\%$ of the unit cell. The basis functions within the muffin-tin spheres are linear combinations of radial wavefunctions and their energy derivatives, computed at energies appropriate to their site and principal, as well as orbital, atomic quantum numbers, whereas outside the muffin-tin spheres the basis functions are combinations of Neuman or Hankel functions^{23,24}. **b**, Energy difference between the stishovite, CaCl_2 , $Pnc2$, $Pa3$, α - PbO_2 and $I2/a$ crystal structures for SiO_2 as a function of volume. The $Pnc2$ structure is used as the zero energy reference level.

calculations at pressures >70 GPa, the density of the *Pnc2* structure becomes greater than that of stishovite; at 100 GPa and 300 K the difference amounts to 0.9% and increases with pressure up to 1.8% at 150 GPa. The *Pnc2* structure becomes more stable than the CaCl_2 -like structure at pressures above 93 GPa and 300 K which agrees well with the FPLMTO (full potential linear muffin-tin orbital; see Fig. 3 legend) result where the transition pressure at 0 K is 80 GPa (as discussed below). We find the following sequence of phase transitions; stishovite $\rightarrow \text{CaCl}_2 \rightarrow \text{Pnc2}$ structure $\rightarrow \text{Pa3}$. The first transition takes place at 45 GPa, the second at 80 GPa and the third at 220 GPa. The transition at 45 GPa is in agreement with the calculation by Kingma *et al.*¹⁰. We have found that the *Pa3* structure is stable at pressures above 220 GPa whereas Cohen²⁰ found the *Pa3* structure to be stable already at 157 GPa because he did not consider the new crystal phase. The transition pressure to the *Pnc2* structure, derived from the present zero-temperature calculations (80 GPa), is somewhat lower than that obtained from the room-temperature lattice dynamics¹⁴ (93 GPa) and molecular dynamics¹⁵ (120 GPa) simulations but higher than the experimental result. A possible reason for this small discrepancy between the predictions of theory might be (1) the difference in theoretical methods and/or (2) the influence of the vibrational energy which is not considered in the first-principles calculations. We thus conclude, based on first-principles calculations, that a *Pnc2* structure must be viewed as a strong candidate for explaining the experimentally observed X-ray results of this study. In Table 2 we present theoretically optimized structural parameters for the *Pnc2* structure at two different volumes.

Because the only experimental data we have available so far is for quenched samples at high pressures, we do not know how the density of SiO_2 would increase (owing to its phase transition) along the geothermal gradient in the mantle. However, the demonstrated reversibility of the phase transition shows clearly that the *Pnc2* phase is the densest phase; on release of pressure the *Pnc2* structure reverses to stishovite (or CaCl_2 -type silica). The only constraints we have on the chemical composition of the lower mantle are from the seismic density data. To reproduce this density profile mineralogically, any combination of silicates and oxides may be chosen. It is therefore important that we know the structure and physical properties of any phase which could be a candidate for the mantle. Through this study, we have established that silica with the *Pnc2* structure is such a phase. $\square$

Received 14 October 1996; accepted 3 June 1997.

1. Sato, H. *et al.* Baddeleyite-type high-pressure phase of TiO_2 . *Science* **251**, 786–788 (1991).
2. Ming, L. S. & Manghnani, M. N. in *High Pressure Research in Geophysics* (eds Manghnani, M. H. & Akimoto, S.) 348–357 (Centre of Academic Publications, Tokyo, 1982).
3. Tang, J. & Endo, S. in *High-Pressure Science and Technology Pt 1* (eds Schmidt, S., Shaner, J. W., Samara, G. A. & Ross, M. K.) 367–370 (Am. Inst. Phys. Woodbury, New York, 1994).
4. Altshuler, L. V. & Produrec, M. A. High-density fluorite and rutile polymorphs. *Solid State Phys.* **15**, 1436–1441 (1973) (in Russian).
5. Simakov, T. New data on compressibility of oxides and fluorites and the theory of homogeneous Earth's composition. *Trans. USSR Acad. Sci.* **310**, 1447–1449 (1991) (in Russian).
6. Sekine, T., Akaishi, M. & Setaka, N. Fe_2N -type SiO_2 from shocked quartz. *Geochim. Cosmochim. Acta* **51**, 379–381 (1987).
7. Liu, L. *High Pressure Research in Geophysics* (eds Manghnani, M. H. & Akimoto, S.) 329–347 (Center of Academic Publications, Tokyo, 1982).
8. Tsuchida, Y. & Yagi, T. A new, post-stishovite high-pressure polymorph of silica. *Nature* **340**, 217–220 (1989).
9. Hemley, R. J. in *High-Pressure Research in Mineral Physics* (eds Syono, Y. & Manghnani, M. H.) 347–359 (Terra Scientific, Tokyo & Am. Geophys. Un., Washington DC, 1987).
10. Kingma, K. J., Cohen, R. E., Hemley, R. J. & Mao, H. K. Transformation of stishovite to a denser phase at lower-mantle pressures. *Nature* **374**, 243–245 (1995).
11. Kingma, K. J., Mao, H. K. & Hemley, R. J. Synchrotron x-ray diffraction of SiO_2 to multimegabar pressures. *High Press. Res.* **14**, 363–374 (1996).
12. Tse, J. S., Klug, D. D. & Page, Y. L. Novel high-pressure phase of silica. *Phys. Rev. Lett.* **69**, 3647–3649 (1992).
13. Lacks, D. J. & Gordon, R. G. Calculations of pressure-induced phase transitions in silica. *J. Geophys. Res.* **98**, 22147–22155 (1993).
14. Dubrovinsky, L. S., Belonoshko, A. B., Dubrovinsky, N. A. & Saxena, S. K. *New High-pressure Silica Phases Obtained by Computer Simulation* 921–923 (AIRAPT/EHPRG Rep., World Scientific, Warsaw, Poland, 1996).
15. Belonoshko, A. B., Dubrovinsky, L. S. & Dubrovinsky, N. A. A new high-pressure silica phase obtained by molecular dynamics. *Am. Mineral.* **81**, 785–788 (1996).
16. Karki, B. B., Warren, M. C., Stixrude, L., Ackland, G. J. & Crain, J. Ab initio studies of high-pressure structural transformations in silica. *Phys. Rev. B* **55**, 3465–3471 (1997).

17. Saxena, S. K., Shen, G. & Lazor, P. Experimental evidence for new iron phase and implications for Earth's core. *Science* **260**, 1312–1314 (1993).
18. Saxena, S. K., Chatterjee, N., Fei, Y. & Shen, G. *Thermodynamic Data on Oxides and Silicates* (Springer, Berlin, 1993).
19. Ross, N. L., Shu, J. F., Hazen, R. M. & Gasparik, T. High-pressure crystal chemistry of stishovite. *Am. Mineral.* **75**, 739–747 (1990).
20. Cohen, R. E. in *High Pressure Research: Application to Earth and Planetary Sciences* (eds Syono, Y. & Manghnani, M. H.) 425–431 (Geophys. Monogr. 67, Am. Geophys. Un., Washington DC, 1992).
21. Willis, J. M. & Cooper, R. B. Synthesis of band and model Hamiltonian theory for hybridizing cerium systems. *Phys. Rev. B* **36**, 3809–3823 (1987).
22. Hedin, L. & Lundqvist, B. I. Explicit local exchange-correlation potentials. *J. Phys. C* **4**, 2064–2083 (1971).
23. Andersen, O. K. Linear methods in band theory. *Phys. Rev. B* **12**, 3060–3083 (1975).
24. Skriver, H. L. *The LMTO Method* (Springer, Berlin, 1984).

Acknowledgements. We thank J. Hu and G. Shen for their help in conducting the X-ray experiments, and N. Dubrovinsky for discussions. This work has been supported by the Swedish Natural Science Research Council (NFR), the Royal Swedish Academy of Sciences, and the Swedish Materials Consortium No. 9 financed by NUTEK, NFR and the Wallenberg Foundation.

Correspondence and requests for materials should be addressed to S.K.S. (e-mail: surendra.saxena@geo.uu.se).

The age of the Popigai impact event and its relation to events at the Eocene/Oligocene boundary

Richard Bottomley*, Richard Grieve†, Derek York‡ & Victor Masaitis§

* Canadian Union College, College Heights, Alberta, T4L 2E5, Canada

† Geological Survey of Canada, Ottawa, Ontario, K1A 0Y3, Canada

‡ Department of Physics, University of Toronto, Toronto, Ontario, M5S 1E7, Canada

§ Karpinsky Geological Institute, St Petersburg, Russia 199026

Ages ranging from the Late Cretaceous (~ 65 Myr) to the Oligocene (~ 29 Myr) have been reported for the 100-km-diameter Popigai impact structure^{1,2} on the Anabar shield, central Siberia. These ages^{3–8} overlap the timing of several possible impact-related features, including the Cretaceous/Tertiary and Eocene/Oligocene stratigraphic boundaries, the North American tektites⁹, and the recently reported occurrences of an iridium anomaly and shocked quartz in Late Eocene deposits in northern Italy¹⁰. Here we report age determinations of several Popigai impact melt rocks using the ^{40}Ar – ^{39}Ar step heating technique to constrain the age of the impact event. Our results are consistent with a Late Eocene impact age of 35.7 ± 0.2 Myr (2σ)—coincident in time with the impact deposits found in Italy. As this age is also similar to that of the North American tektites, which have been associated with the Chesapeake Bay impact structure in the eastern United States^{11–13}, there seem to have been at least two large and essentially contemporaneous impacts during the Late Eocene.

The target rocks at the Popigai structure (111°E , $71^\circ 30'\text{N}$) are Precambrian, generally granitic, gneisses overlain by ~ 1.25 km of sandstones and carbonates, ranging in age from Proterozoic to Cretaceous^{1,2}. The impact-related lithologies are capped by several metres of Pliocene and Quaternary alluvial and glacial deposits. In addition to voluminous allochthonous breccias, Popigai contains $\sim 2,000$ km³ of impact melt rocks, including glass-bearing suevite breccias, and so-called tagamites, the latter occurring as irregular melt bodies and coherent sheets, up to several hundred metres thick¹.

The stratigraphic age of Popigai is between the end of the Cretaceous (~ 65 Myr) and the beginning of the Pliocene (~ 5 Myr). Dating impact events using isotopic methods is often not straightforward^{14,15} and previous attempts to date the Popigai impact have given age estimates that range from Late Cretaceous to Oligocene. Earlier whole-rock K–Ar age estimates include: 40–45 Myr (mean 42.5 Myr; ref. 3) for four tagamites and 29–46 Myr (mean 38.7 Myr; ref. 4) for five tagamites and one glass fragment

from a suevite breccia. Fission track analyses of six glass fragments from suevite breccia yielded 32.4–44.0 Myr (mean 38.9 Myr; ref. 1). An additional fission track analysis gave 30.5 ± 1.2 Myr (ref. 5). Deino *et al.*⁶ reported five total fusion ^{40}Ar – ^{39}Ar ages ranging from 52.2 to 65.9 Myr on glass from suevite breccia, and plateau ages of 65.5 ± 5 Myr and 65.4 ± 0.3 Myr, that is, a Cretaceous/Tertiary (K/T) age. Most recently, Deino *et al.*⁷ failed to reproduce a K/T age from additional samples. They obtained a highly discordant ^{40}Ar – ^{39}Ar spectrum from a fused gneiss fragment and an integrated age of 37.7 ± 0.3 Myr from four impact melt glasses, which they considered to be a maximum for the Popigai impact event. An age for Popigai of 39 ± 9 Myr, consistent with both the whole rock K–Ar and most of the fission track ages, has been generally quoted⁸.

We have undertaken a stepwise heating ^{40}Ar – ^{39}Ar analyses of six tagamite samples from Popigai in an attempt to resolve the uncertainty in age and the possible relationship to various events in the stratigraphic record. Our previous experience with ^{40}Ar – ^{39}Ar has shown that analyses from a suite of well re-equilibrated impact melt rocks gives the least ambiguous estimate of the true age of impact¹⁵. The tagamite (impact melt rock) samples analysed here range from glassy (P72, P22), through cryptocrystalline (sub-microscopic crystals; P1, P84), hypocrySTALLINE (crystals in a glassy ground-mass; P21) to holocrystalline (all crystals; P9). Paralleling the increasing matrix crystallinity, there is a reduction in clasts from

~20% to <10%. The clasts are dominantly quartz and feldspar and, in the relatively coarser-grained varieties, have indications of partial melting and reaction with the matrix, with the development of so-called chequerboard feldspar and orthopyroxene-high silica glass coronas around quartz^{16,17}. Where identifiable, the matrix mineralogy is hypersthene (40–60% ferrosilite molecule) and labradorite (55–65% anorthite molecule) laths in a ground-mass of glass or an intergrowth of quartz and potassium feldspar. The major-element chemistry of the samples is similar to previous analyses of Popigai tagamites². Trace-element data indicate variable, but elevated, siderophile abundances compared to the Archaean target rock gneisses. Iridium and nickel abundances in the samples analysed here range from 0.54 to 1.7 p.p.b. and 77 to 112 p.p.m., respectively, and it has been suggested that Popigai was formed by the impact of an ordinary chondrite body¹⁸.

The ^{40}Ar – ^{39}Ar age spectra have a variety of shapes, which reflect the complex underlying argon systematics in the tagamites. We show the spectra from the least to the most disturbed. Replicate runs are denoted by the letters A and B following the sample identifier. The age spectra from holocrystalline sample P9 and hypocrySTALLINE sample P21 show the classic pattern of mild disturbance in the lower-temperature fractions, where argon loss from the less retentive sites lowers the age of the early fractions. As gas is released from the more retentive sites at higher temperatures, the ages stabilize around a

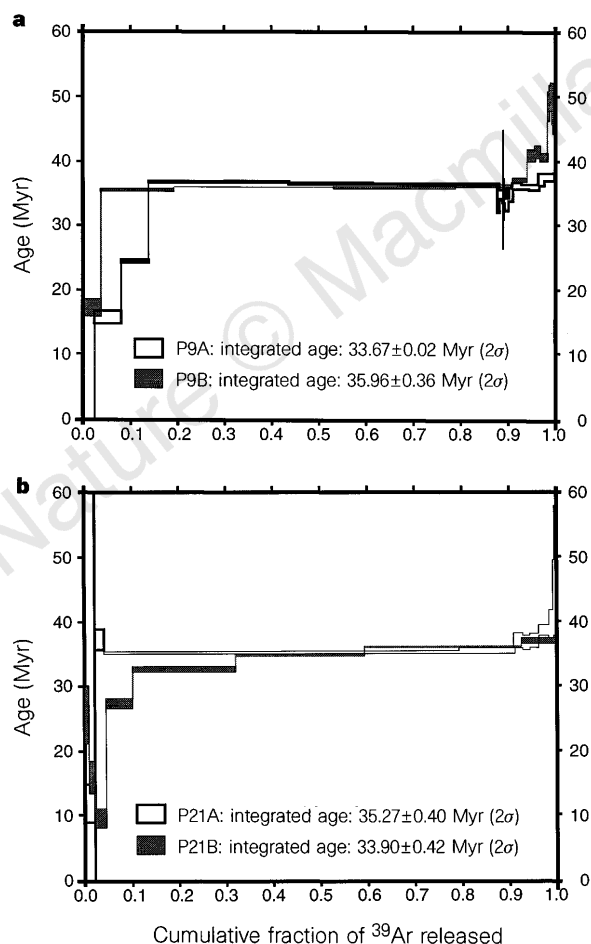


Figure 1 ^{40}Ar – ^{39}Ar incremental heating spectra for holocrystalline (P9; **a**) and hypocrySTALLINE (P21; **b**) impact melt rocks from Popigai. The vertical width of each fraction represents the experimental error for that fraction at the 1σ level. The integrated or total gas ages for each run (corresponding to conventional K–Ar ages) are also indicated for each sample. The standard used for these determinations was hb3gr with an age of 1,071 Myr (refs 21, 22). Details of the experimental data are available; see Supplementary Information.

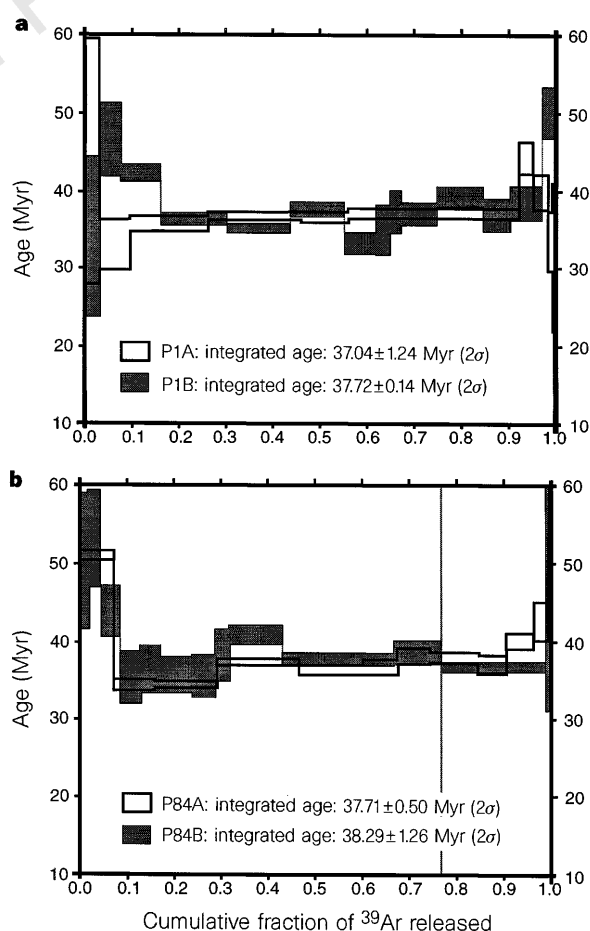


Figure 2 ^{40}Ar – ^{39}Ar incremental heating spectra for cryptocrystalline impact melt rocks from Popigai. **a**, P1; **b**, P84.

plateau that typically involves at least 80% of the gas released (Fig. 1a, b). Spectrum 21A has a plateau age of 35.7 ± 0.2 Myr (2σ), whereas spectra from 9A (plateau, 36.7 ± 0.2 Myr) and 9B (plateau, 36.3 ± 0.2 Myr) give a slightly older apparent age. The replicate tagamite chip in run 21B has lost a larger percentage of its argon in the lower fractions and displays a stepwise rise throughout most of its release, never reaching a true plateau.

The cryptocrystalline samples P1 and P84 display more structured spectra (Fig. 2a, b), where the individual fraction ages fluctuate around a common age but with more scatter than expected from the experimental errors (indicated by the vertical thickness of the box for each fraction on the figure). Whereas P1A shows a good plateau, with 84% of its release indicating an age of 36.9 ± 0.6 Myr, the spectra of the replicate, P1B, oscillates around this value. Both P84A and P84B oscillate around means of 35.6 ± 1.2 Myr and 37.5 ± 0.8 Myr, but with inter-fraction differences ranging up to 5 Myr.

Spectra from the glassy samples P72A and P72B (Fig. 3a) superficially resemble structured plateaux and are the result of two different phases releasing the bulk of their argon at different temperatures. The phase preferentially releasing its gas at lower temperatures is characterized by a low apparent Ca/K ratio (as indicated by the ratio of $^{37}\text{Ar}/^{39}\text{Ar}$). The higher-temperature phase, dominating the last 50% of the gas released, has a Ca/K ratio that is

30% higher. The integrated ages of 37.7 ± 0.6 and 35.8 ± 1.8 Myr are broadly concordant with the ages of the better-behaved tagamite samples, which give plateaux. This is consistent with sample P72 having undergone an internal redistribution of argon between mineral phases without loss, presumably due to ^{39}Ar recoil during neutron irradiation¹⁹. Spectra P22A and P22B (Fig. 3b) display broad U-shaped spectra, which are characteristic of contamination by clasts partially degassed by the event that reset the argon clock²⁰. In samples displaying this shape, the minimum in the trough of the U gives an upper limit to the age of the resetting event. The minimum age is 52 Myr for sample P22A and 42 Myr for P22B, indicating an upper limit for the impact event of 42 Myr.

As the formation of impact melt involves the melting of target rocks and the incorporation of clastic debris, argon ages can range from the age of impact up to the age of the target rock, depending on how well the argon clock was reset in the various components of the melt rock. This suite of tagamite samples displays such an array of argon behaviour, as evidenced by their varied spectral shapes. There is a correlation between the degree of structure in the spectra and the petrographic character of the tagamite samples. The best behaved argon spectra come from the rocks which show greater matrix crystallinity and degree of thermal equilibration of the relict mineral inclusions (for example, P9, P21). Non-plateau spectra come from the tagamites with a higher proportion of glass and unequilibrated inclusions. In a situation where we are dealing with contamination by older material and some modification of the argon clock—either by differential mineralogy or thermal resetting of lower-retentivity sites—we consider the best estimate of the age of impact is given by the sample which displays the best developed plateau and youngest age: sample P21A. Using the four fractions which make up the plateau portion of P21A, we find an age of impact of 35.7 ± 0.2 Myr (2σ). For comparison, the average of the plateau ages of the next best behaved samples, P9A and P9B, is 36.5 ± 0.4 Myr (2σ), which we believe reflects some relict argon contamination. This age (35.7 Myr) is broadly consistent with earlier reported Late Eocene age determinations, using conventional K/Ar and fission track dating, but the resolution and interpretative techniques possible with ^{40}Ar – ^{39}Ar dating allow an improved determination of the time of impact. In comparing age determinations of contemporaneous events, it should be noted that there is an uncertainty in the age of the standard (monitor) of the order of 1% (refs 21,22), and measures of laboratory precision do not always fully reflect the underlying geological complexity.

There is no evidence to support the involvement of Popigai in the K/T event (65 Myr). The difference in our determination of the age of Popigai (35.7 ± 0.2 Myr) and the Eocene/Oligocene boundary (33.7 ± 0.5 Myr)²³ would also preclude a role for Popigai. The Late Eocene stratigraphic record, however, contains at least two Ir anomalies¹⁰ and deposits marked by microspherules and microtektites^{9,24}, as well as shocked quartz and coesite^{25,26}. In the Eocene/Oligocene stratotype section at Massignano, Italy, a distinct Ir peak occurs at the lower calcareous nanofossil NP19/20 zone (mid-lower foraminiferal P16 zone, top of chron 16n2). Based on dating by various isotopic methods of volcanic ash in the same section, the age of this Ir anomaly is 35.7 ± 0.4 Myr (ref. 10). The Ir anomaly is associated with shocked quartz containing planar deformation features^{27,28}. Given the almost identical age for Popigai and that it has been suggested¹⁸ that it was caused by the impact of a chondritic projectile, the Popigai impact event would seem a logical source for both the Ir and the shocked quartz in the Late Eocene section in northern Italy.

There is a second, smaller Ir anomaly in the same section within magnetic chron 16r1, some 0.2 Myr earlier¹⁰. In addition, an Ir anomaly associated with microspherules and microtektites is known from Barbados and several Deep Sea Drilling Program cores in the Late Eocene²⁵. All, or some of these, are related to the North American tektite strewn field. Unfortunately, uncertain

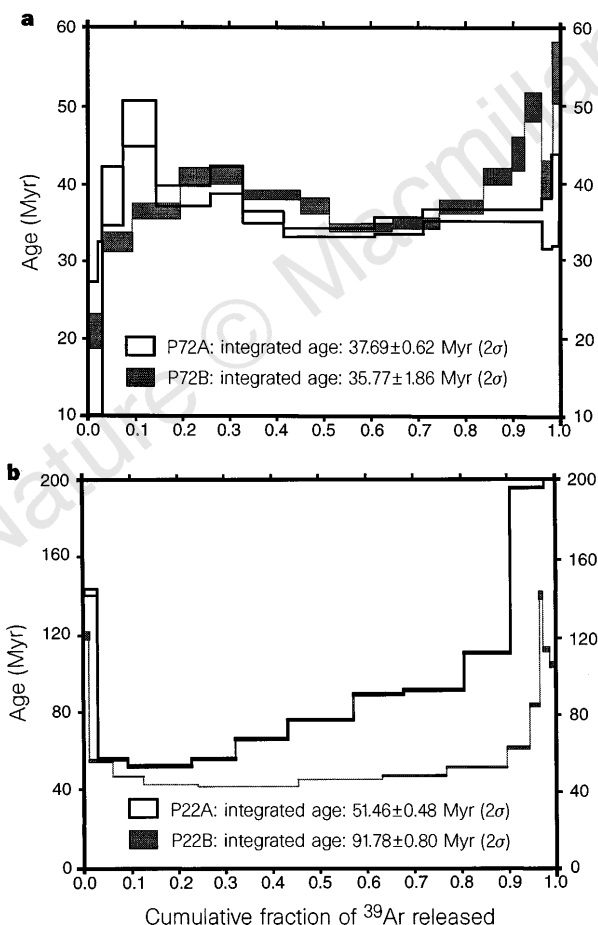


Figure 3 ^{40}Ar – ^{39}Ar incremental heating spectra for glassy impact melt rocks from Popigai. **a**, P72; **b**, P22.

biostratigraphic correlations precludes a determination of whether they are strictly coeval¹⁰. ⁴⁰Ar–³⁹Ar plateau age determinations of tektite fragments in DSDP 612 give 35.5 ± 0.3 Myr (ref. 29), indistinguishable from 35.4 ± 0.6 Myr for tektites from Barbados³⁰. It is highly unlikely that Popigai is also the source of the North American tektites, as isotopic evidence, as well as geographical location, tends to exclude Popigai as a source³¹. The best candidate seems to be the recently discovered 85-km-diameter Chesapeake Bay impact structure¹¹ ($76^\circ 5' \text{ W}$, $37^\circ 15' \text{ N}$), which formed during the early part of the foraminiferal P15 zone and the calcareous nanofossil NP 19/20, that is, at 35.5–35.2 Myr (ref. 12). The Chesapeake Bay structure has the advantage of having much more suitable target rocks¹³ and location close to the occurrence of North American tektites.

Isotopic dating can not prove that the Ir and shocked quartz in the Massignano section are from Popigai or that the North American tektites are from Chesapeake Bay. This will require the discovery of appropriate geochemical and/or mineralogical markers from the source crater in the corresponding distal ejecta. It is clear, however, that two large impact events resulting in 100- and 85-km diameter structures occurred within, at most, a few hundred thousand years of each other in the Late Eocene. In terms of the terrestrial cratering rate, this represents an unusually small time between events; the expected³² repeat rate of impact structures of this size is of the order of ten million years. □

Received 13 December 1996; accepted 5 June 1997.

- Masaitis, V. L., Mikhailov, M. V. & Selivanovskaya, T. V. *The Popigai Meteorite Crater*. (Nauka, Moscow, 1975). (in Russian.)
- Masaitis, V. L. *et al.* *The Geology of Astroblemes* (Nedra, Leningrad, 1980). (in Russian.)
- Firsov, L. V. Paleogene basaltoids in the Popigai graben, Anabar Shield. *Dokl. Akad. Nauk SSSR* **194**, 664–666 (1970). (in Russian.)
- Komarov, A. N. & Raykhlin, A. I. Comparison of fission-track and potassium-argon dating of impactites. *Dokl. Akad. Nauk SSSR* **228**, 673–676 (1976).
- Storzer, D. & Wagner, G. A. Fission track dating of El'gygytgyn, Popigai and Zhamanshin impact craters: No sources for Australasian or North-American tektites. (abstr.) *Meteoritics* **14**, 541–542 (1979).
- Deino, A. L., Garvin, J. B. & Montanari, S. K/T age for the Popigai impact event? (abstr.) *Lunar Planet. Sci.* **XXII**, 297–298 (1991).
- Deino, A. L., Montanari, A. & Vishnevsky, S. in *4th Int. Workshop ESF Network "Impact Cratering and Evolution of Planet Earth"* (eds Montanari, A. & Coccioni, R.) 66–67 (Ancona, Italy, 1995).
- Grieve, R. A. F. Terrestrial impact structures. *Annu. Rev. Earth Planet. Sci.* **15**, 245–270 (1987).
- Glass, B. P. & Zwart, M. J. in *Stratigraphic Micropaleontology of Atlantic Basin and Boderlands* (ed. Swain, F. M.) 553–568 (Elsevier, New York, 1979).
- Montanari, A., Asaro, E., Michel, H. V. & Kennett, J. P. Iridium anomalies of Late Eocene age at Massignano (Italy), and ODP Site 689B (Maud Rise, Antarctic). *Palaio* **8**, 420–437 (1993).
- Poag, C. W., Powars, D. S., Poppe, L. J. & Mixon, R. B. Meteoroid mayhem in Ole Virginny: Source of the North American tektite strewn field. *Geology* **22**, 691–694 (1994).
- Poag, C. W. & Aubry, M.-P. Upper Eocene impactites of the U.S. East Coast: Depositional origins, biostratigraphic framework, and correlation. *Palaio* **10**, 16–43 (1995).
- Stecher, O., Ngo, H. H., Papanastassiou, D. A. & Wasserburg, G. J. Nd and Sr isotopic evidence for the origin of tektite materials from DSDP site 612 off the New Jersey coast. *Meteoritics* **24**, 89–98 (1989).
- Deutsch, A. & Schärer, U. Dating terrestrial impact events. *Meteoritics* **29**, 301–322 (1994).
- Bottomley, R. J., York, D. & Grieve, R. A. F. ⁴⁰Argon–³⁹Argon dating of impact craters. *Proc. Lunar Planet. Sci. Conf.* **20**, 421–431 (1990).
- Grieve, R. A. F. Petrology and chemistry of the impact melt at Mistastin Lake crater, Labrador. *Geol. Soc. Am. Bull.* **86**, 1617–1629 (1975).
- Floran, R. J. *et al.* Manicouagan impact melt, Quebec, 1, stratigraphy, petrology, and chemistry. *J. Geophys. Res.* **83**, 2737–2759 (1978).
- Masaitis, V. L. & Raikhlin, A. I. The Popigai crater formed by the impact of an ordinary chondrite. *Dokl. Akad. Nauk SSSR* **286**, 1476–1478 (1986). (in Russian.)
- Turner, G. & Cadogan, P. H. Possible effects of ³⁹Ar recoil in ⁴⁰Ar–³⁹Ar dating. *Proc. Lunar Sci. Conf.* **5**, 1601–1615 (1974).
- Harrison, T. M. & McDougall, I. Excess ⁴⁰Ar in metamorphic rocks from Broken Hill, New South Wales; implications for ⁴⁰Ar/³⁹Ar age spectra and the thermal history of the region. *Earth Planet. Sci. Lett.* **55**, 123–149 (1981).
- Layer, P. W., Hall, C. M. & York, D. The derivation of ⁴⁰Ar–³⁹Ar age spectra of single grains of hornblende and biotite by laser step-heating. *Geophys. Res. Lett.* **14**, 757–760 (1987).
- Turner, G., Huneke, J. C., Podosek, F. A. & Wasserburg, G. J. ⁴⁰Ar–³⁹Ar ages and cosmic ray exposure age of Apollo 14 samples. *Earth Planet. Sci. Lett.* **12**, 19–35 (1971).
- Montanari, A. Geochronology of the terminal Eocene impacts: An update. *Geol. Soc. Am. Spec. Pap.* **274**, 607–616 (1990).
- Glass, B. P. & Burns, C. A. Late Eocene crystal-bearing spherules: Two layers or one? *Meteoritics* **22**, 265–279 (1987).
- Glass, B. P. North American tektite debris and impact ejecta from DSDP site 612. *Meteoritics* **34**, 209–218 (1989).
- Bohor, B. F., Betterton, W. J. & Foord, E. E. Coesite, glass and shocked quartz at DSDP Site 612: Evidence for nearby impact in the late Eocene. (abstr.) *Lunar Planet. Sci.* **XIX**, 114–115 (1988).
- Clymer, A. K., Bice, D. M. & Montanari, A. Shocked quartz from the late Eocene: Impact evidence from Massignano, Italy. *Geology* **24**, 483–486 (1996).
- Langenhorst, F. Characteristics of shocked quartz in late Eocene impact ejecta from Massignano (Ancona, Italy): Clues to shock conditions and source crater. *Geology* **24**, 487–490 (1996).
- Obradovich, J. D., Snee, L. W. & Izett, G. A. Is there more than one glassy impact layer in the late Eocene? (abstr.) *Geol. Soc. Am. Abstr. Progr.* **2**, A134 (1989).

- Glass, B. P., Hall, C. M. & York, D. ⁴⁰Ar/³⁹Ar laser-probe dating of North American tektite fragments from Barbados and the age of the Eocene-Oligocene boundary. *Chem. Geol.* **59**, 181–186 (1986).
- Shaw, H. F. & Wasserburg, G. J. Age and provenance of the target materials for tektites and possible impactites as inferred from Sm–Nd and Rb–Sr systematics. *Earth Planet. Sci. Lett.* **60**, 155–177 (1982).
- Grieve, R. A. F. & Shoemaker, E. M. in *Hazards Due to Comets and Asteroids* (ed. Gehrels, T.) 417–462 (Univ. Arizona Press, Tucson, 1994).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from Mary Sheehan at the London editorial office of Nature.

Acknowledgements. We thank G. B. Dalrymple, S. Kelley, M. Pilkington and M. Villeneuve for comments on earlier version of this Letter. This work was supported by an NSERC operating grant to D.Y.

Correspondence should be addressed to R.G. (e-mail: impact@gsc.nrcan.gc.ca).

Earliest known Old World monkey skull

Brenda R. Benefit & Monte L. McCrossin

Department of Anthropology, Southern Illinois University, Carbondale, Illinois 62901, USA

Similarities of the skull are commonly used to support hypotheses of ancestor–descendant relationships between fossil and living ape genera, especially between the late Miocene apes *Sivapithecus* and *Dryopithecus* from Eurasia and the living orang-utan (*Pongo*) from Borneo and Sumatra^{1–4}. Yet determining whether craniofacial traits shared by extant and Miocene apes are primitive or derived is severely hampered by the rarity of well-preserved fossil crania, particularly of early members of their closest outgroup, the Old World monkeys (Cercopithecoidea). The discovery of a complete and undistorted skull of *Victoriapithecus* at middle Miocene deposits from Maboko Island, Kenya, provides evidence of intact cranial-vault and basicranial morphology, brain size and craniofacial hafting for a primate from between 32 and 7 million years ago. *Victoriapithecus* represents a branch of Old World monkey that is intermediate between extant cercopithecids (Colobinae and Cercopithecinae) and the common ancestor they shared with apes (Hominoidea)^{5–8}. The skull preserves traits widely thought to be derived for extant and fossil members of a proposed *Sivapithecus/Pongo* clade, but which now appear to be primitive features of ancestral Old World higher primates in general.

Victoriapithecus cranium KNM-MB 29100 was found *in situ* during excavation of sediment at Maboko Island that has been radiometrically dated as older than 14.7 Myr (ref. 9), and biostratigraphically dated as younger than 16 Myr (ref. 10). The cranium has preserved left male canine and postcanine teeth on both sides (Fig. 1, Table 1). Anatomy of the teeth and skull show *Victoriapithecus* to be a catarrhine (Old World higher primate) of modern aspect, and more specifically an Old World monkey (Fig. 2). The absence of at least 30 derived dental features shared by the living Old World monkey subfamilies Colobinae and Cercopithecinae demonstrates that *Victoriapithecus* belongs to neither extant group, but represents the sister taxon (Victoriapithecidae) of modern Cercopithecoidea^{5–8}.

The cranial vault of *Victoriapithecus* differs from those of all extant Old World monkeys in being narrower and lower relative to length (Table 1), a configuration found only in the early Pliocene African colobine *Libypithecus*¹¹. *Victoriapithecus* and *Libypithecus* also share unusually well-developed sagittal and nuchal crests. Relative to a body-weight estimate of 4.5 kg for large male postcrania¹², the cranial capacity of KNM-MB 29100 at 54 cm³ falls between regressions for extant anthropoids and strepsirrhines, and is significantly higher than that of the Oligocene archaic catarrhine *Aegyptopithecus*¹³ (Fig. 3). The relative brain size of *Victoriapithecus* is more similar to that of the red colobus than to any other living cercopithecoid¹⁴.

The craniofacial hafting and upper facial morphology of *Victoriapithecus* are unique among Old World monkeys, adding to the list of traits which demonstrate that victoriapithecids represent

the sister taxon of modern cercopithecids. The face is only slightly ventrally deflected (klinorhynch) relative to the basicranium. Its profile is steep and unusually linear, with the bridge of the nose lying along a straight line connecting prosthion, rhinion and glabella. This linear profile is combined with a slight dorsal orientation of the orbits and zygomatic frontal process, with their inferior borders positioned well anterior to their superior margins. Of the living catarrhines, the upper facial morphology of *Victoriapithecus* is most similar to that of *Pongo* (the orang-utan), in having tall and narrow orbits (Table 1), well-developed supraorbital costae, which together with anteriorly convergent temporal lines and absence of a postglabellar depression contribute to the formation of a frontal trigon, and presence of three zygomatic foramina positioned well above the inferior orbital rim on the frontal process of the zygomatic. In addition, the cheek region of the zygomatic is tall relative to facial height, and its root extends only a short distance above the molar alveolae, as in Miocene apes but not modern catarrhines¹⁵.

The skull confirms previous assessments based on isolated facial bones^{6,15} that aspects of the face of *Victoriapithecus* resemble those of modern cercopithecines in having a narrow interorbital septum, low position of the frontozygomatic suture, narrow nasal bones, a low and narrow nasal aperture, a moderately long and anteriorly tapering snout, and a moderately long premaxilla with upper first incisor (I¹) positioned anterior to upper second incisor (I²) (Table 1). Many of these features are adaptations for frugivory¹⁶, a diet indicated for *Victoriapithecus* by dental evidence⁷, and which is characteristic of cercopithecine monkeys. Because parsimony indicates that morphology shared by victoriapithecids and either or both cercopithecid subfamilies should be primitive for Cercopithecoidea^{7,12,15,17}, a predominantly cercopithecine-like cranium and frugivorous diet characterized the earliest Old World monkeys. This position is confirmed by strong cranial resemblances between Victoriapithecidae and the basal catarrhine outgroup Propliopithecidae^{18,19}. The fossil colobine *Libypithecus*¹¹ retains

Figure 1 Anterior (a), posterior (b), left lateral (c), superior (d) and inferior (e) views of KNM-MB 29100.

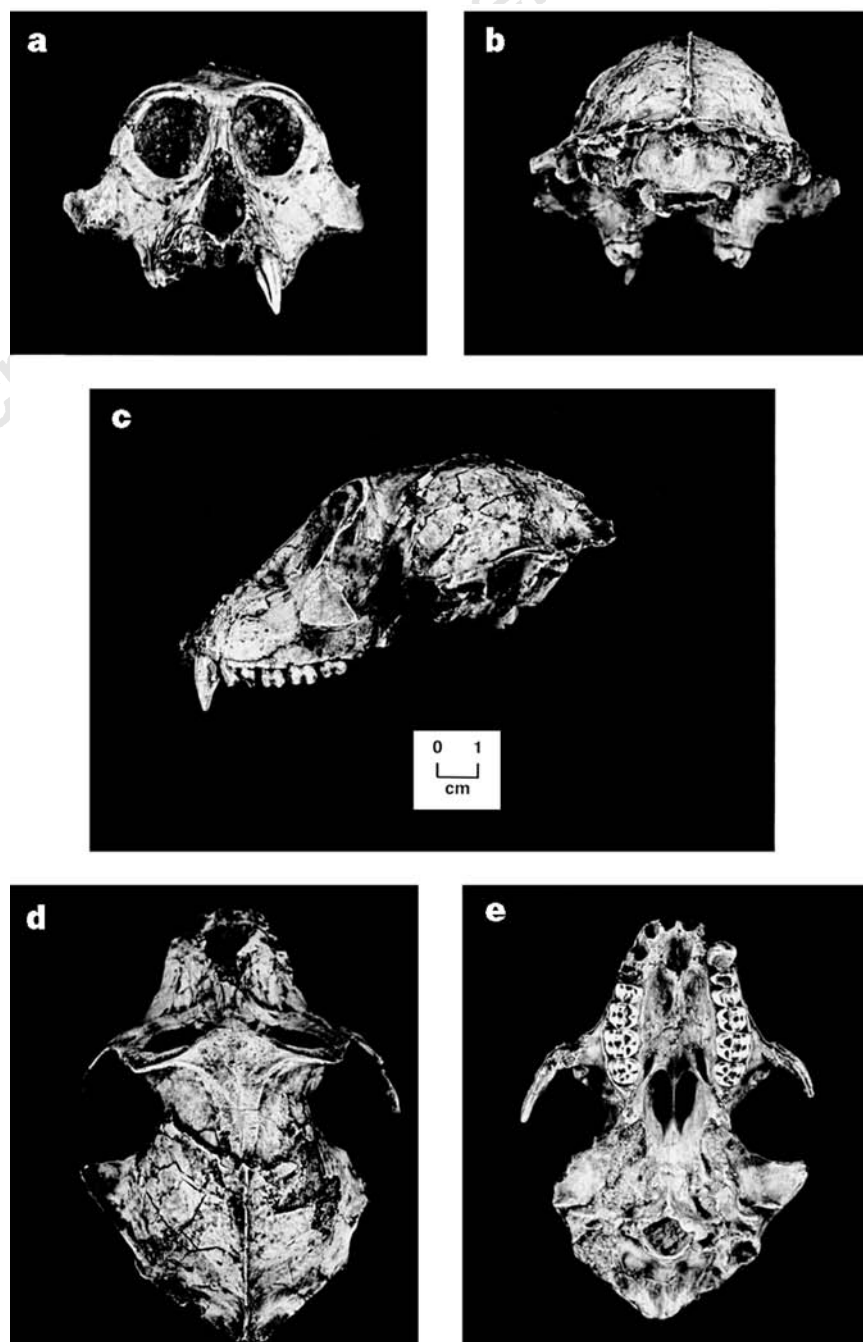


Table 1 Cranial measurements (mm) and comparative indices (%) of KNM-MB 29100

	KNM-MB 29100	<i>Libypithecus</i>	Extant colobine	Extant cercopithecine	<i>Pongo</i>	Other extant apes	<i>Sivapithecus</i>	<i>Aegyptopithecus</i>
Cranial vault								
Height (basion to bregma)	42.4							
Width (euryon to euryon)	43.6							
Length (nasion to inion)	75.5							
Postglenoid process to distal M ³	40							
Width × 100/length	58	55	70–74	71–79	85	67–78		72
Height × 100/length	56	53	58–62	59–63	86	61–82		47
Height × 100/width	97	97	82–86	84–89	101	85–84		70
Postglenoid process to M ³ × 100/width	92	96	55–66	56–78	79	58–83		61
Orbit								
Biorbital breadth (internal)	41.6							
Interorbital breadth	6.9							
Orbit height (maximum)	21.2							
Orbit width (maximum)	17.4							
Interorbital breadth × 100/biorbital breadth	16	14	16–22	11–16	16	21–30	12	25
Orbital width × 100/orbit height	85	94	100–107	98–119	85	98–102	75	92
Snout								
Malar height (frontomaxillary sutures)	16.6							
Facial height (orbitale to molar alveolae)	23.6							
Zygomatic height (above alveolus)	7.2							
Snout length projected horizontally	35.6							
Face length (orbit to nasal aperture)	19.8							
Palate length (distal M ² –prosthion)	46							
Palate width (externally across M ² s)	39							
Palate depth (maximum)	7.9							
Premaxilla length (suture to prosthion)	10.9							
Nasal aperture width (maximum)	12							
Snout length × 100/palate length	77		46–68	60–93	70	49–71	78	66
Facial length × 100/facial height	84	83	8–47	35–86	26	21–42		84

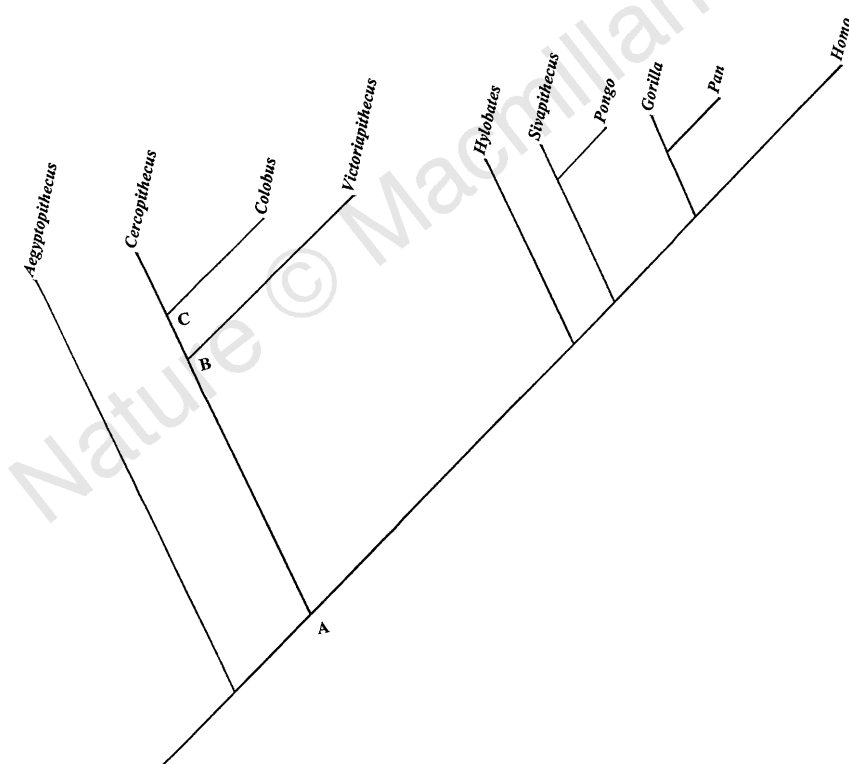


Figure 2 *Victoriapithecus* shares a tubular ectotympanic with modern catarrhines^{17,22} (node A). Features linking the *Victoriapithecus* skull with Old World monkeys include continuation of the mesial sulcus onto the C¹ root and absence of a maxillary sinus^{715–1722,23} (node B). Teeth of living cercopithecines and colobines share several derived features not found in *Victoriapithecus*, including: dp^{3–4} and M^{1–3} possessing a distal transverse loph, lacking a *crista obliqua*, and being longer than wide; well-developed dp₃ hypoconid, entoconid and transverse protolophid; presence of dp₄ hypolophid; absence of dp₄ and M_{1–2} hypoconulids; P₄ aligned with the molar row; and relatively longer M₁ (refs 5, 7, 8) (node C).

much of this primitive morphology, including a moderately long snout, although its teeth are adapted for folivory. Other cercopithecids, and especially modern colobines, are comparatively derived.

Because many primitive cercopithecoid cranial features shared by *Victoriapithecus* and the propliopithecid *Aegyptopithecus* are also found among Miocene hominoids (especially *Afropithecus*^{20,21}), they are probably primitive characteristics for Old World higher primates in general. Such primitive features include low cranial vault height, a well-developed sagittal crest, a frontal trigon created by distinct supraorbital costae and anteriorly convergent temporal lines, slight ventral deflection of the face, a linear facial profile,

tall and narrow orbits, a moderately long snout, and a relatively tall cheek region. Three of these primitive catarrhine features (supraorbital costae, a frontal trigon, and tall cheek regions) are present in the facial skeletons of both *Sivapithecus*^{1,2} and *Dryopithecus*^{3,4}, and tall and narrow orbits are characteristic of the former.

The new evidence refutes the idea that ancestral Old World anthropoids had high and rounded braincases and short faces, as reconstructed from presumed conservative resemblances between colobines and gibbons^{16,17,22,23}. The skull morphology of *Victoriapithecus* demonstrates that many aspects of the craniofacial anatomy of late Miocene large-bodied hominoids are primitive

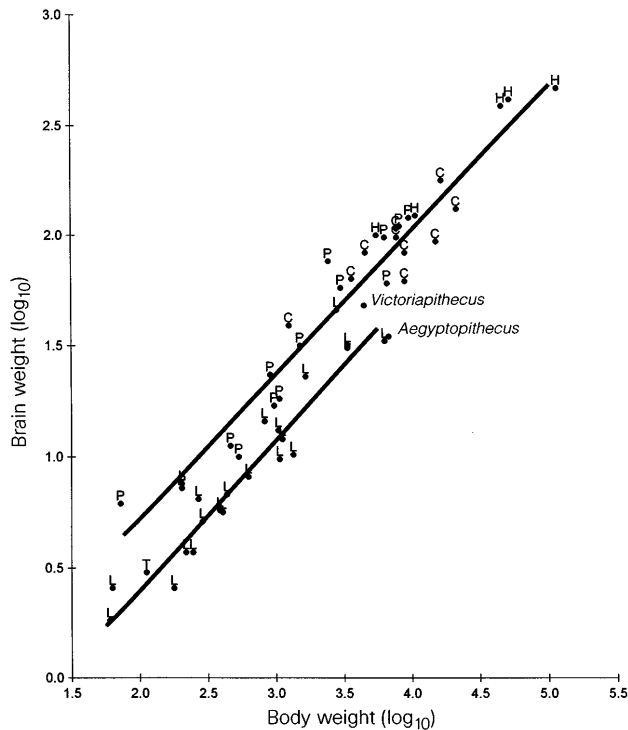


Figure 3 Bivariate plot of $\log_{10}$ -transformed mean brain and body weight data for extant hominoids (H), cercopithecoidea (C), platyrrhines (P), tarsiers (T), lemuroids and lorisoidea (L)¹⁴, *Victoriapithecus* and *Aegyptopithecus*. Superimposed best fit lines for extant anthropoids (higher) and strepsirrhines (lower) are based on least-squares linear regression equations.

catarrhine features, rather than derived indicators of affinity with the great ape and human clade. In this way, the *Victoriapithecus* skull shows that the anatomy of fossil cercopithecoidea is as important as that of hominoids for deciphering the evolutionary history of Old World higher primates. □

Received 21 November 1996; accepted 30 May 1997

- Pilbeam, D. R. New hominoid skull material from the Miocene of Pakistan. *Nature* **295**, 232–234 (1982).
- Brown, B. & Ward, S. in *Orang-utan Biology* (ed. Schwartz, J. H.) 247–260 (Oxford Univ. Press, New York, 1988).
- Moya-Sola, S. & Kohler, M. Recent discoveries of *Dryopithecus* shed new light on evolution of great apes. *Nature* **365**, 543–545 (1993).
- Moya-Sola, S. & Kohler, M. New partial cranium of *Dryopithecus* Lartet, 1863 (Hominoidea, Primates) from the upper Miocene of Can Llobateres, Barcelona, Spain. *J. Hum. Evol.* **29**, 101–139 (1995).
- Von Koenigswald, G. H. R. Miocene Cercopithecoidea and Oreopithecoidea from the Miocene of East Africa. *Foss. Vert. Afr.* **1**, 39–51 (1969).
- Benefit, B. R. & McCrossin, M. L. Ancestral facial morphology of Old World higher primates. *Proc. Natl Acad. Sci. USA* **88**, 5267–5271 (1991).
- Benefit, B. R. The permanent dentition and phylogenetic position of *Victoriapithecus* from Maboko Island, Kenya. *J. Hum. Evol.* **25**, 83–172 (1993).
- Benefit, B. R. Phylogenetic, paleodemographic and taphonomic implications of *Victoriapithecus* deciduous teeth from Maboko, Kenya. *Am. J. Phys. Anthropol.* **95**, 277–331 (1994).
- Feibel, C. S. & Brown, F. H. Age of the primate-bearing deposits on Maboko Island, Kenya. *J. Hum. Evol.* **21**, 221–225 (1991).
- Andrews, P., Meyer, G., Pilbeam, D. R., Van Couvering, J. A. & Van Couvering, J. A. H. The Miocene fossil beds of Maboko Island, Kenya: Geology, age, taphonomy and paleontology. *J. Hum. Evol.* **10**, 35–48 (1981).
- Stromer, E. Mitteilungen über die Wirbeltierreste aus dem Mittelpliocen des Natrontales (Ägypten). 1. Affen. *Z. Deutsch. Geol. Gesell. Abh.* **65**, 349–361 (1913).
- Harrison, T. New postcranial remains of *Victoriapithecus* from the middle Miocene of Kenya. *J. Hum. Evol.* **18**, 3–54 (1989).
- Radinsky, L. The fossil evidence of anthropoid brain evolution. *Am. J. Phys. Anthropol.* **41**, 15–28 (1974).
- Martin, R. D. *Primate Origins and Evolution* (Princeton Univ. Press, Princeton, NJ, 1990).
- Benefit, B. R. & McCrossin, M. L. The facial anatomy of *Victoriapithecus* and its relevance to the ancestral cranial morphology of Old World monkeys and apes. *Am. J. Phys. Anthropol.* **92**, 329–370 (1993).
- Vogel, C. Morphologische Studien am Gesichtsschädel Catarrhiner Primaten. *Biblioth. Primatol.* **4**, 1–226 (1966).
- Harrison, T. The phylogenetic relationships of the early catarrhine primates: A review of the current evidence. *J. Hum. Evol.* **16**, 41–80 (1987).
- Simons, E. L. The earliest apes. *Sci. Am.* **217**, 28–35 (1967).
- Simons, E. L. New faces of *Aegyptopithecus* from the Oligocene of Egypt. *J. Hum. Evol.* **16**, 273–289 (1987).

- Leakey, R. E. & Leakey, M. G. A new Miocene hominoid from Kenya. *Nature* **342**, 143–146 (1986).
- Leakey, M. G., Leakey, R. E., Richtsmeier, J. T., Simons, E. L. & Walker, A. C. Similarities in *Aegyptopithecus* and *Afropithecus* facial morphology. *Folia Primatol.* **56**, 65–85 (1991).
- Delson, E. & Andrews, P. in *Phylogeny of the Primates: A Multidisciplinary Approach* (eds Luckett, W. P. & Szalay, F. S.) 405–446 (Plenum, New York, 1975).
- Strasser, E. & Delson, E. Cladistic analysis of cercopithecoidea relationships. *J. Hum. Evol.* **16**, 81–99 (1987).

Acknowledgements. Excavations at Maboko were conducted with permission of the Office of the President, Republic of Kenya and in collaboration with the National Museums of Kenya. We thank the field crew (especially B. Onyango, V. Oluoch and S. Gitau) and M. G. Leakey for assistance, and E. Delson, M. Kohler, S. Moya-Sola and D. Pilbeam for comments and advice. This work was supported by NSF, L. S. B. Leakey Foundation, National Geographic Society, Wenner Gren Foundation for Anthropological Research, Fulbright, and Boise Fund.

Correspondence and requests for materials should be addressed to B.R.B. (e-mail: bbenefit@siu.edu).

Apparent competition structures ecological assemblages

M. B. Bonsall & M. P. Hassell

Department of Biology and the NERC Centre for Population Biology, Imperial College at Silwood Park, Ascot, Berkshire SL5 7PY, UK

Competition is a major force in structuring ecological communities¹. It acts directly² or indirectly, in which case it may be mediated by shared natural enemies and is known as ‘apparent competition’^{3–6}. The effects of apparent competition on species coexistence are well known theoretically^{7,8} but have not previously been demonstrated empirically in controlled multigenerational experiments. Here we report on the population dynamic consequences of apparent competition in a laboratory insect system with two host species and a common parasitoid attacking them. We find that whereas the two separate, single host–single parasitoid interactions are persistent, the three-species system with the parasitoid attacking both hosts species (which are not allowed to compete directly) is unstable, and that one of the host species is eliminated from the interaction owing to the effects of apparent competition.

Classical ecological theory predicts that simple interactions in which species share common natural enemies are unstable, leading to one species being eliminated from the interaction⁹. This effect arises through competitive interactions mediated by the natural enemy in ‘apparent competition’^{7,8}. As the two host species are not directly competing for resources, the loss of one of the species results solely from the dynamic consequences of the natural enemy becoming more abundant as a consequence of having an alternative host species. The species that persists in the interaction is the one that can support the higher parasitoid density⁶.

These indirect competitive interactions have received little attention from natural or laboratory systems, although they are likely to be as important, if not more so, than direct competitive interactions. Empirical work so far has focused mainly on short-term, behaviourally orientated studies in which natural enemy responses are recorded within a single field season of the interaction^{10–15}. No empirical studies have addressed directly the long-term effects of apparent competition on the population dynamics of the interacting organisms. We now report on the temporal consequences of such apparent competition (Figs 1, 2). A series of long-term laboratory experiments were used to explore the interaction between the ichneumonid parasitic wasp *Venturia canescens* (Gravenhorst) and two of its moth hosts, *Plodia interpunctella* (Hubner) and *Ephestia kuehniella* Zeller, over a period of several generations. The direct competitive effects between the host species were fully excluded by a vertical barrier of nylon mesh dividing each cage in half. The mesh size was large enough to allow free access for the searching parasitoids between the two

Figure 1 Representative examples of the population dynamics of the single host-single parasitoid interactions: **a, b**, *P. interpunctella* (dotted line)-*V. canescens* (dashed line); **c, d**, *E. kuehniella* (unbroken line)-*V. canescens* (dashed line). Experiments were carried in cages (30 × 30 × 30 cm) under standard environmental conditions (25 ± 2°C; 70 ± 5% R.H. 16:8 h light-dark cycle). Linear time series analysis indicates an underlying tendency for oscillations dampening towards a stable equilibrium.

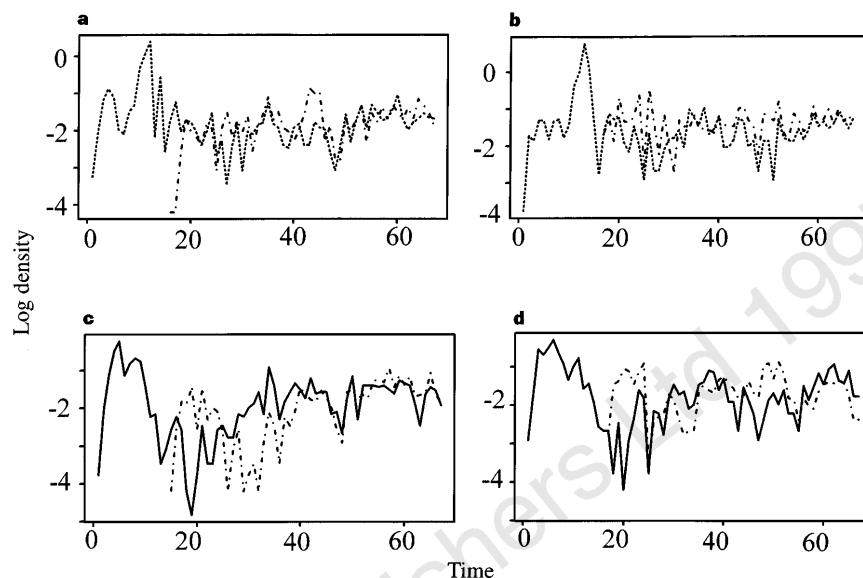
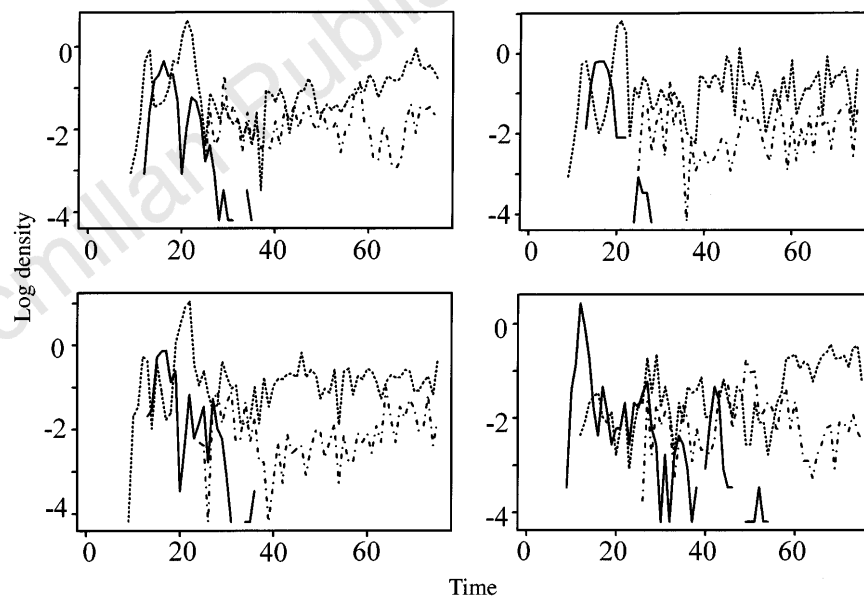


Figure 2 Representative examples of the population dynamics of the three species system of two hosts and a single parasitoid: *P. interpunctella* (dotted line)-*E. kuehniella* (unbroken line)-*V. canescens* (dashed line). In all these cases, *E. kuehniella* is eliminated from the interaction by the action of the shared natural enemy, *V. canescens*. Time series analysis now indicates damped oscillations for the interaction between *P. interpunctella* and *V. canescens* and diverging oscillations for *E. kuehniella*.



halves, but was too small to let the hosts pass through. The interactions with a single host and single parasitoid, and the more complex three-species interactions of two hosts and a single parasitoid, were each replicated eight times and monitored over 15 generations by weekly census counts of dead adults¹⁶.

In all the two-species systems, the two populations persisted in relatively stable interactions (Fig. 1). Time-series statistics show that the dynamics in these cases were largely driven by delayed density-dependent processes (primarily due to the parasitoids) and predict damped oscillations to a stable equilibrium. In contrast, the full three-species interaction of *P. interpunctella*-*E. kuehniella*-*V. canescens* never persisted (Fig. 2), and in all cases *E. kuehniella* was always rapidly eliminated from the system by the action of the shared natural enemy. Time-series statistics again show that the population dynamics of the three species were governed by delayed density dependence. But now *E. kuehniella* is predicted to show divergent oscillations while *P. interpunctella* and *V. canescens* show damped oscillations towards a stable equilibrium. The parasitoid *V. canescens* thus directly restricts the number of host species that it attacks, a phenomenon known as 'dynamic monophagy'⁶.

The results reveal interesting properties of the interactions. A repeated-measures analysis of variance shows that the interaction

between the two apparently competing host species (*P. interpunctella* and *E. kuehniella*) is amensal: the effect of *P. interpunctella* on *E. kuehniella* is strong whereas the reciprocal effects of *E. kuehniella* on *P. interpunctella* is negligible (Fig. 3). *P. interpunctella* is dominant because it has a larger intrinsic rate of increase and shorter development time¹⁶.

Such asymmetric relationships in the interaction strengths have been widely documented for direct competition² and also recorded in one case for an indirect interaction¹⁷. If shown to be normal for indirect competitive interactions as well, amensal apparent competition may prove to be a key process for structuring ecological assemblages. It could, for example, play a central part in the persistence of species assemblages made up of a single dominant species (free from the effects of direct and apparent competition) and many rarer species reduced in abundance by the effects of indirect amensal interactions.

We have demonstrated empirically the long-term population effects of two species sharing a natural enemy. The resulting extinction of one of the host species mimics direct competition, although direct competitive effects are absent. Such indirect effects could be very important for the dynamics of species assemblages if they are widespread under natural conditions. Host-parasitoid

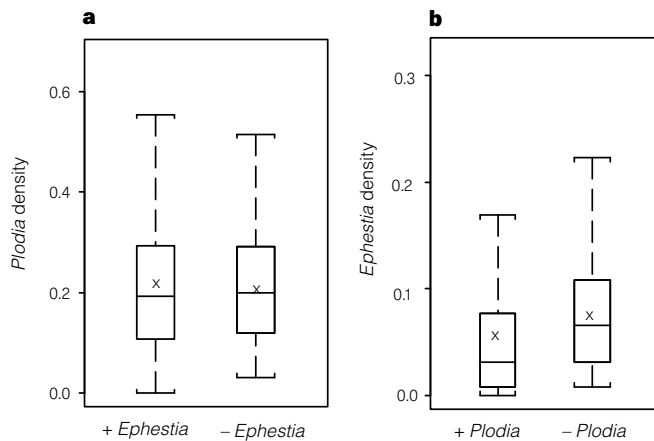


Figure 3 Box plots for **a**, *E. kuehniella*, and **b**, *P. interpunctella* in the presence (+) or absence (–) of the other host species. The solid line within each box marks the median; the cross marks the mean value. The length of the box is the interquartile range and the limit bars show the distance to 1.5 times the interquartile range giving a measure of dispersion in the data. The plots show the density of each species over a nine-week period. A repeated-measures analysis of variance (following a square-root transformation) reveals that the interaction mediated via the shared parasitoid, *V. canescens*, is amensal. The effects of apparent competition are extremely skewed, with one host suffering very little while the other is driven to extinction.

assemblages may be particularly prone to these indirect effects (compare with other natural enemy–victim systems) because parasitoids tend to have a pronounced numerical response and to limit host population sizes well below their carrying capacities¹⁸. Generalist predators, on the other hand, will often show weaker numerical responses to particular prey species and therefore be less likely to cause species exclusion through apparent competition¹⁹. □

Received 17 March; accepted 28 April 1997.

1. Law, R. & Watkinson, A. R. in *Ecological Concepts* (ed. Cherratt, J. M.) (Blackwell Scientific, London, 1989).
2. Lawton, J. H. & Hassell, M. P. Asymmetrical competition in insects. *Nature* **289**, 793–795 (1981).
3. Menge, B. A. Indirect effects in marine rocky intertidal interaction webs—patterns and processes. *Ecol. Monogr.* **65**, 21–74 (1995).
4. Strauss, S. Y. Indirect effects in community ecology: their definition, study and importance. *Trends Ecol. Evol.* **6**, 206–210 (1991).
5. Wootton, J. T. The nature and consequences of indirect effects in ecological communities. *Annu. Rev. Ecol. Sys.* **25**, 443–466 (1994).
6. Holt, R. D. & Lawton, J. H. Apparent competition and enemy-free space in insect host–parasitoid communities. *Am. Nat.* **142**, 623–645 (1993).
7. Holt, R. D. Predation, apparent competition and the structure of prey communities. *Theor. Pop. Biol.* **12**, 197–229 (1977).
8. Holt, R. D. Spatial heterogeneity, indirect interactions and the coexistence of prey species. *Am. Nat.* **124**, 377–406 (1984).
9. Williamson, M. H. An elementary theory of interspecific competition. *Nature* **180**, 422–425 (1957).
10. Müller, C. B. & Godfray, H. C. J. Apparent competition between two aphid species. *J. Anim. Ecol.* **66**, 57–64 (1997).
11. Karban, R., Houghen-Eitzmann, D. & English-Loeb, G. Predator-mediated apparent competition between two herbivores that feed on grapevines. *Oecologia* **97**, 508–511 (1994).
12. Settle, W. H. & Wilson, L. T. Invasion by the variegated leafhopper and biotic interactions: parasitism, competition and apparent competition. *Ecology* **71**, 1461–1470 (1990).
13. Schmitt, R. J. Indirect interactions between prey: apparent competition, predator aggregation and habitat segregation. *Ecology* **68**, 1887–1897 (1987).
14. Hanley, K. A., Vollmer, D. M. & Case, T. J. The distribution and prevalence of helminths, coccidia and blood parasites in two competing species of gecko: implications for apparent competition. *Oecologia* **102**, 220–229 (1994).
15. Grosholz, E. D. Interactions of intraspecific, interspecific and apparent competition with host–pathogen dynamics. *Ecology* **73**, 507–514 (1992).
16. Bonsall, M. B. *Temporal and Spatial Insect Population Dynamics* Thesis, Univ. London (1997).
17. Evans, E. W. & England, S. Indirect interactions in biological control of insects: pests and natural enemies in alfalfa. *Ecol. Appl.* **6**, 920–930 (1996).
18. Beddington, J. R., Free, C. A. & Lawton, J. H. Characteristics of successful natural enemies in models of biological control of insect pests. *Nature* **273**, 513–519 (1978).
19. Holt, R. D. in *Multitrophic Interactions in Terrestrial Systems* (eds Gange, A. C. & Brown, V. K.) (Blackwell Scientific, London, 1996).

Acknowledgements. We thank R. Holt for comments on an earlier version of the manuscript. This work was supported by an NERC studentship to M.B.B. and an NERC-funded research grant to M.P.H.

Correspondence and requests for materials to be addressed to M.B.B. (e-mail: m.bonsall@ic.ac.uk).

Genetic interaction between male mating strategy and sex ratio in a marine isopod

Stephen M. Shuster* & Clay Sassaman†

* Department of Biological Sciences, Northern Arizona University, Flagstaff, Arizona 86011-5640, USA

† Department of Biology, University of California, Riverside, California 92521, USA

Individual males in many animal species exhibit discrete modes of behaviour^{1–3}, but the genetic mechanisms underlying these differences are poorly understood. Here we investigate the genetics of the isopod crustacean *Paracerceis sculpta*, in which three different types of males coexist, each distinguishable from the others by their behavioural and morphological phenotypes^{4,5}. Within families, alleles of the gene encoding the enzyme phosphoglucosyltransferase (*Pgm* gene) are associated with particular male phenotypes, although no significant association between these characters exists population-wide. This suggests that *Pgm* is closely linked to a single genetic locus which controls male phenotype. We call this the *alternative mating strategy* (*Ams*) locus. We present evidence that two other factors—an autosomal gene, *transformer* (*Tfr*), and an extrachromosomal factor—interact with primary sex determination loci and with alleles at *Ams*, causing certain individuals to change sex, thereby biasing family sex ratios. A model based on our genetic analysis suggests that: first, polymorphism in male behaviour is controlled by the mendelian segregation of three alleles at the *Ams* locus; second, that family sex ratio is influenced by alternative alleles at the *Tfr* locus whose expression is influenced by the extrachromosomal factor; and third, that *Tfr* and *Ams* interact epistatically to determine the sex of the individual and, if male, its behaviour and external morphology.

Females are monomorphic in *P. sculpta*. Males, however, exhibit three distinct morphs that differ in reproductive behaviour: α -males are largest and defend harems within sponges using elongated posterior appendages; β -males invade harems by mimicking female behaviour and morphology; and γ -males invade harems by being small and secretive^{5,6}. A genetic model has been proposed⁷ to explain the persistence of the three male morphs at stable frequencies, in which three alleles at a single autosomal locus (*Ams*) show directional dominance and mendelian inheritance.

We tested this model⁷ for male morphology using controlled laboratory crosses. We first examined mendelian inheritance at the *Pgm* locus⁸ in 25/31 F₁ families in which one heterozygous and one homozygous parent were crossed. As three alleles were detectable at *Pgm*, we summarized alleles possessed by heterozygous parents as allele 1 or allele 2. The total numbers of progeny possessing these allele classes were 405 and 424, respectively, and individual crosses were homogeneous (*G*-test (ref. 9), *G*_H = 31.01) (d.f. = 24, *P* > 0.10). We considered *Pgm* inheritance to be mendelian.

The genetic model⁷ suggested that field-collected β - and γ -males are heterozygous at the *Ams* locus, and thus should produce 50:50 ratios of α - and β -, or α - and γ -male sons, respectively, when crossed with field-collected females (see Methods). As most β - and γ -males were also heterozygous at the *Pgm* locus (11/13 and 7/10, respectively), we examined the association between *Pgm* genotype and male phenotype among F₁ progeny. If these loci were unlinked, we expected the progeny of males heterozygous at both loci to segregate four combinations of the two male morphs and two *Pgm* genotypes in equal frequency. The more severe the deviation from this expectation, the closer the linkage, thus, parental and recombinant classes were each pooled, then compared using a *G*-test (d.f. = 1)⁹.

Of the 351 male progeny reared from 18 double-heterozygote *Pgm*–*Ams* crosses, 96.3% appeared in the parental class ($G = 368.9$, $P < 0.001$), a result indicating that *Ams* is closely linked to *Pgm* (separated by four map units), and one consistent with the hypothesis that a major gene causes phenotypic differences among males⁷. We confirmed, moreover, that male differences are not simply due

to allelic differences at *Pgm*. Electrophoretic analysis over a two-year period (1987–88) showed no evidence of linkage disequilibrium⁹ between *Pgm* allelomorphs and the three male morphs in nature ($D^2 = 0.004$, $P > 0.10$, $N = 292$). Thus *Pgm* and *Ams*, although linked, represent distinct genetic loci.

To examine mendelian inheritance at *Ams*, we unambiguously

Table 1 *Paracerceis sculpta* F₁ crosses

Cross-class	No. of families	No. of progeny	Weighted survivorship	Progeny phenotypes					Expected male frequency			G_{Ams}	Proportion of males†	$G_{sex\ ratio}$
				α	β	γ	F	N	α	β	γ			
<i>Ams</i> ^a <i>Ams</i> ^a × <i>Ams</i> ^a <i>Ams</i> ^a	8	247	0.43	53	0	0	55	108	1.0:0.00:0.00			0.00	0.49	0.37
<i>Ams</i> ^b <i>Ams</i> ^a × <i>Ams</i> ^a <i>Ams</i> ^a	12	1,308	0.49	59	267	0	317	643	0.50:0.50:0.00			143.60**	0.51	0.13
<i>Ams</i> ^b <i>Ams</i> ^a × <i>Ams</i> ^b <i>Ams</i> ^a	1	107	0.39	0	28	0	14	42	0.25:0.75:0.00			<0.001‡	0.67	4.75*
<i>Ams</i> ^a <i>Ams</i> ^a × <i>Ams</i> ^b <i>Ams</i> ^a	10	921	0.38	75	0	105	167	347	0.50:0.00:0.50			5.02*	0.52	0.49
	31	2,583	0.44	187	295	105	553	1,140						

F_1 , number of females; N , total progeny; $G_{Ams} = G_{adj}$ (d.f. = 1) comparison of observed and expected male frequencies; $G_{sex\ ratio} = G_{adj}$ (d.f. = 1) comparison of observed and 1:1 sex ratio.

* $P < 0.05$; ** $P < 0.001$.

† Proportion of males weighted by family size.

‡ Exact χ^2 test.

Table 2 Distribution of *Pgm* alleles among F₁ males and females in cross-classes 2 and 4

Cross-class	P ₁ Q ₁	P ₁ Q ₂	P ₂ Q ₁	P ₂ Q ₂	N	G_P	G_Q	G_{PQ}	Inference
<i>Ams</i> ^a <i>Ams</i> ^a × <i>Ams</i> ^a <i>Ams</i> ^a	27	4	0	12	43	8.69*	2.85	33.00*	Sex change (1x)
	33	12	0	29	74	3.49	0.87	36.99*	Sex change (2x)
	10	1	2	13	26	0.62	0.15	17.45*	Sex change (2x)
	18	7	0	17	42	1.53	0.86	20.38*	Sex change (2x)
	19	11	0	3	33	25.64*	0.76	3.18	Sex change (1x)†
	14	0	10	11	35	1.41	4.95*	6.44*	Sex change (1x)†
	10	1	9	18	38	6.95*	0.00	8.88*	Sex change (1x)
	16	0	14	26	56	10.63*	0.29	14.65*	Sex change (1x)
	27	1	0	19	47	1.73	1.05	55.48*	Sex change (2x)
	17	1	0	10	28	2.32	1.30	30.19*	Sex change (2x)
<i>Ams</i> ^a <i>Ams</i> ^a × <i>Ams</i> ^b <i>Ams</i> ^a	4	7	0	0	11	**	0.83	0.83	Sex change (1x)†
	5	5	6	9	25	1.01	0.36	0.36	Independent assort
	4	6	13	15	38	8.88*	0.42	0.00	Sex change (1x)†
	1	5	7	6	19	2.64	0.48	1.33	Independent assort
	10	10	8	9	37	0.24	0.03	0.03	Independent assort
	14	9	1	5	29	4.33*	0.47	5.94*	Sex change (1x)
	3	7	12	7	29	2.84	0.03	2.84	Independent assort

P₁, sex; Q₁, *Pgm* genotype; thus, P₁Q₁ represents β -males with *Pgm* allele class 1; P₁Q₂ are α -males with *Pgm* allele class 2; P₂Q₁ are females bearing a β -allele as indicated by *Pgm* allele class 1; P₂Q₂ are females bearing an α -allele as indicated by *Pgm* allele class 2. Inferences are explained in Methods. * $P < 0.05$; ** $P < 0.001$.

† Crosses that appear to have undergone 1x or 2x sex change, but whose pattern of G_P , G_Q and G_{PQ} tests were not consistent with confirmed sex-changed crosses because *Pgm* data were unavailable for all progeny; however, the identity of these crosses were confirmed with exact χ^2 tests in Table 3a.

Table 3 Inheritance of *Ams*, *Tfr* and ECF in progeny of *P. sculpta*

Ams cross-type	Tfr cross-type	ECF state	No. of families	No. of progeny	Weighted survivorship	Observed				Expected				N	Exact χ^2 probability	
						α -	β -	γ -	F	α -	β -	γ -	F			
(a) Among F ₁ progeny																
Ams ^a Ams ^a × Ams ^a Ams ^a	Tfr ¹ Tfr ¹ × Tfr ¹ Tfr ¹	M(-); F(-)	1	85	0.34	15	0	0	14	14.5	0	0	14.5	29	1.00	
Ams ^b Ams ^a × Ams ^a Ams ^a	Tfr ¹ Tfr ² × Tfr ² Tfr ²	M(+); F(-)	2	194	0.72	0	37	0	87	0	45.88	0	78.12	124	0.11	
	Tfr ¹ Tfr ² × Tfr ¹ Tfr ²	M(+); F(-)	1	95	0.49	5	13	0	29	2.82	20.68	0	23.5	47	0.06	
			1	115	0.41	3	9	0	35	2.82	20.68	0	23.5	47	0.00*	
	Tfr ¹ Tfr ¹ × Tfr ² Tfr ²	M(+); F(-)	2	227	0.40	0	51	0	41	0	46	0	46	92	0.35	
	Tfr ¹ Tfr ¹ × Tfr ¹ Tfr ²	M(+); F(-)	5	589	0.49	41	134	0	118	38.09	146.5	0	108.41	293	0.34	
	Tfr ¹ Tfr ¹ × Tfr ¹ Tfr ¹	M(+); F(-)	1	85	0.47	10	23	0	7	10	20	0	10	40	0.55	
	Tfr ¹ Tfr ¹ × Tfr ² Tfr ²	M(+); F(-)	1	107	0.39	0	28	0	14	0	31.5	0	10.5	42	0.28	
	Ams ^b Ams ^a × Ams ^a Ams ^a	Tfr ¹ Tfr ² × Tfr ² Tfr ²	M(-); F(-)	1	93	0.16	1	0	4	10	1.8	0	3.75	9.45	15	0.87
	Ams ^a Ams ^a × Ams ^a Ams ^a	Tfr ¹ Tfr ² × Tfr ² Tfr ²	M(-); F(-)	6	579	0.39	47	0	47	132	56.5	0	56.5	113	226	0.05
				1	94	0.36	5	0	11	18	8.5	0	8.5	17	34	0.02*
	Tfr ¹ Tfr ¹ × Tfr ¹ Tfr ¹	M(-); F(-)	2	155	0.46	22	0	43	7	18	0	36	18	72	0.55	
Totals			24	2,418	0.43	149	295	105	512	153.03	331.24	104.75	471.98	1,061		
(b) Among F ₂ progeny																
Ams ^a Ams ^a × Ams ^a Ams ^a	Tfr ¹ Tfr ² × Tfr ² Tfr ²	M(-); F(+)	2	52	0.58	0	0	0	30	0	0	0	30	30	1.00	
	Tfr ¹ Tfr ² × Tfr ¹ Tfr ²	M(-); F(+)	1	29	0.31	3	0	0	6	1.17	0	0	7.83	20	0.10	
Ams ^a Ams ^a × Ams ^b Ams ^a	Tfr ¹ Tfr ² × Tfr ² Tfr ²	M(-); F(+)	1	60	0.45	0	14	0	13	0	9.99	0	17.01	27	0.17	
	Tfr ¹ Tfr ¹ × Tfr ² Tfr ²	M(-); F(+)	2	78	0.26	0	11	0	9	0	10	0	10	20	0.82	
Ams ^b Ams ^a × Ams ^a Ams ^a	Tfr ¹ Tfr ² × Tfr ² Tfr ²	M(+); F(+)	4	148	0.28	0	16	0	25	0	15.17	0	25.83	41	0.87	
	Tfr ¹ Tfr ² × Tfr ¹ Tfr ²	M(+); F(+)	3	109	0.40	4	16	0	24	2.64	19.36	0	22	44	0.46	
	Tfr ¹ Tfr ¹ × Tfr ² Tfr ²	M(+); F(+)	1	37	0.62	0	12	0	11	0	11.5	0	11.5	23	1.00	
	Tfr ¹ Tfr ¹ × Tfr ¹ Tfr ²	M(+); F(+)	4	158	0.34	8	23	0	23	7.02	27	0	19.98	54	0.57	
Ams ^b Ams ^a × Ams ^b Ams ^a	Tfr ¹ Tfr ¹ × Tfr ² Tfr ²	M(+); F(+)	1	46	0.57	0	15	0	11	0	14.62	0	11.38	26	1.00	
Ams ^b Ams ^a × Ams ^a Ams ^a	Tfr ¹ Tfr ² × Tfr ² Tfr ²	M(-); F(-)	2	63	0.35	0	8	8	6	0	8.14	4.18	9.68	22	0.08	
Ams ^a Ams ^a × Ams ^a Ams ^a	Tfr ¹ Tfr ² × Tfr ¹ Tfr ²	M(-); F(+)	1	36	0.39	0	0	6	8	1.82	0	3.5	8.68	14	0.15	
	Tfr ¹ Tfr ² × Tfr ² Tfr ²	M(-); F(+)	1	44	0.43	1	0	7	11	4.75	0	4.75	9.5	19	0.13	
	Tfr ¹ Tfr ² × Tfr ² Tfr ²	M(-); F(+)	1	40	0.65	0	15	0	11	0	9.62	3.38	13	26	0.04*	
	Tfr ¹ Tfr ¹ × Tfr ¹ Tfr ²	M(-); F(+)	1	35	0.40	0	6	0	8	0	7	1.82	5.18	14	0.23	
Totals			25	935	0.40	16	136	21	196	25.98	132.4	17.63	192.99	369		

Ams, alternative mating strategy locus; *Tfr*, transformer locus; ECF, extrachromosomal factor; M, male; F, female; F , number of females; N , total number of progeny; +, positive for ECF; -, negative for ECF. * $P < 0.05$.

assigned parental males and females to four cross-classes using *Pgm* genotypes in phase with α -, β - and γ -male phenotypes (Table 1). We pooled the total progeny within cross-classes to obtain the observed frequencies of females and males of different phenotypes. We then compared these frequencies with those predicted by the genetic model⁷. The predicted phenotypes appeared in the four cross-classes (Table 1), a result consistent with the presumed dominance relationships among the *Ams* alleles ($Ams^{\beta} > Ams^{\gamma} > Ams^{\alpha}$)⁷. However, according to strict mendelian expectations, crosses involving β - and γ -sires showed an excess of β - and γ -sons, respectively (Table 1). Moreover, pooled sex ratios deviated from unity only in cross-class 3 (Table 1), whereas family sex ratios within cross-classes 2 and 4 were heterogeneous ($G_H = 62.88$ (d.f. = 9, $P < 0.001$) and 19.45 (d.f. = 5, $P < 0.005$), respectively). Given evidence that *Pgm* and *Ams* are closely linked, apparent mendelian inheritance at *Pgm*, but not at *Ams*, as well as heterogeneous sex ratios within cross-classes, indicated that additional factors must influence male phenotype.

Both genetic and extrachromosomal factors affect family sex ratios in peracarid crustaceans^{10–17}. Certain autosomal genes cause genetic females to mature as males, whereas extrachromosomal factors (from either bacteria or virions) appear to produce the opposite effect^{10–17}. As the *Pgm/Ams* complex in *P. sculpta* is autosomal⁷, we reasoned that if sex-ratio biasing factors exist, they should cause deviations in the observed frequencies of *Pgm* alleles between males and females, in the same families that showed sex-ratio biases and excesses of β - and γ -males. Differently put, if sex-ratio biasing factors caused males to mature as females and females to mature as males, we predicted an apparent interaction between *Pgm* and sex, even though these traits were unlinked (Fig. 1).

		Q1	Q2		G_P	G_Q	G_{PQ}	Inference
a	P1	10	10	20	NS	NS	NS	Independent assortment
	P2	10	10	20				
		20	20	40				
		Q1	Q2					
b	P1	0	10	10	NS	*	NS	Lethal allele
	P2	0	10	10				
		0	20	20				
		Q1	Q2					
c	P1	0	0	0	*	NS	NS	Sex-limited lethal
	P2	10	10	10				
		10	10	20				
		Q1	Q2					
d	P1	20	10	30	*	NS	*	Sex change (1x)
	P2	0	10	10				
		20	20	40				
		Q1	Q2					
e	P1	20	0	20	NS	NS	*	Sex change (2x)
	P2	0	20	20				
		20	20	40				
		Q1	Q2					

Figure 1 Detection of sex-ratio biasing factors in *P. sculpta*. 2×2 tables show hypothetical allelic combinations among 40 progeny produced by a P_1P_2 , Q_1Q_2 male crossed with a P_1P_2 , Q_1Q_1 female; P represents sex (P_1 , males; P_2 , females); Q = *Pgm* (Q_1 , allele class 1; Q_2 , allele class 2); G tests (G_P , G_Q , G_{PQ}) show significant (asterisk) and non-significant (NS) deviations in sex ratio and *Pgm* allele frequency, as well as interactions between sex and *Pgm*; **a**, independent assortment; **b**, lethal allele indicates mortality regardless of sex; **c**, sex-limited lethal indicates mortality regardless of Q_i genotype; **d**, P_2Q_1 interaction causes females to mature as males; **e**, P_2Q_1 and P_1Q_2 interactions cause females to mature as males and males to mature as females.

To test this hypothesis, we examined the distribution of *Pgm* alleles among male and female progeny in F_1 cross-classes 2 ($Ams^{\beta}Ams^{\alpha} \times Ams^{\alpha}Ams^{\alpha}$) and 4 ($Ams^{\gamma}Ams^{\alpha} \times Ams^{\alpha}Ams^{\alpha}$) (Tables 1, 2). Independent assortment occurred in 4/17 F_1 crosses, all involving γ -sires. The remaining 13/17 F_1 crosses showed evidence of one-way or two-way sex change (Table 2), indicating that sex-ratio biasing factors exist in *P. sculpta*. Lethal sex-limited or autosomal alleles are unlikely to have produced these deviations because sex-ratio biases occurred in both directions (G_P in Table 2), and because *Pgm* showed mendelian inheritance in 16/17 crosses (G_Q in Table 2). Variation in male-morph and sex-ratio frequencies, within cross-classes presumed to be homogeneous with respect to their *Ams*- and sex-determination genotypes (for example, cross-classes 2 and 4, in Table 2), indicated that expression of sex-ratio biasing factors is contingent on individuals' allelic states at *Pgm/Ams* and at primary sex-determination loci. Moreover, evidence of two-way sex change suggested that sex-ratio biasing factors in *P. sculpta* were both genetic and extrachromosomal^{10,12,13,15–17}.

We devised a model to explain the observed variation in F_1 phenotype frequencies in which one autosomal (*Tfr*, for transformer) and one extrachromosomal factor (ECF) exist (factor effects are explained in Methods; Fig. 2). Like previous hypotheses regarding sex-ratio biasing factors^{10,12,13,15–17}, we assumed that primary γ sex determination involved female heterogamety⁷. Unlike previous models, however, we assumed that two-way sex change is caused by alternative alleles at a single autosomal locus (*Tfr*), and that the effects of ECF are contingent on allelic states at *Ams*, *Tfr* and primary sex-determination loci (Fig. 2). Our model thus provided a testable explanation for male- and female-biased sex ratios, for one- and two-way sex change, for the inheritance of the *Pgm/Ams*

		<i>Tfr</i> genotype		
<i>Ams</i> genotype	Sex	111	112	212
α/α	M	-	+	+
	F	-	↓*	↓
β/α	M	-	-	-
	F	↑	↑*	-
γ/α	M	-	-	-
	F	↑	-	-

Figure 2 Effects of *Tfr* and ECF on *Ams* and primary sex determination loci: α , β , γ are alleles at *Ams*; M, male; F, female; Tfr^i represent alleles at *Tfr*; minus sign, no effect of *Tfr* on *Ams*-sex combination; plus sign, sex change, with an arrow indicating the direction of change; asterisk, effect produced by interaction of ECF with *Tfr*, *Ams* and primary sex-determination genotypes. The apparent effect of ECF is to enhance or suppress the expression of the Tfr^2 allele depending on allelic states at an individual's *Ams* and primary sex-determination loci.

complex, and for the surplus of β - and γ -males in F_1 families (Table 1; Fig. 3).

We tested our model by assigning *Ams* and *Tfr* genotypes, as well as ECF states (Table 3a) to all parental individuals using unambiguous *Pgm* genotypes and apparent *Tfr* genotypes (Table 2). We then compared observed and expected frequencies of male and female F_1 progeny using exact χ^2 tests¹⁸. We found no significant deviation of observed from expected frequencies in 22/24 F_1 families (for N progeny, $N = 1,061$; Table 3a). Moreover, the presumed *Tfr*-genotypes of parental individuals conformed to Hardy–Weinberg expectations (exact $\chi^2 = 0.21$).

We further tested our model by combining (1) unambiguous *Ams* genotypes (determined from the *Pgm* genotypes of F_1 parents and F_2 families; Tables 1, 2), with (2) predicted *Tfr* genotypes for F_2 progeny (determined from the apparent *Tfr* genotypes of their F_1 parents; Table 2), with (3) the predicted ECF state of F_2 progeny (determined from the apparent ECF state of their F_1 parents; Table 3a), to generate expected male-morph frequencies and sex ratios for all F_2 progeny (Table 3b). Using the same methods for comparing observed and expected progeny frequencies as those described for F_1 families, we found no significant deviation in male morph or sex ratios in 24/25 F_2 families (Table 3b). Although ECF was initially detectable only in parental β -males, this factor was evidently transmitted to individuals of both sexes, because this factor's interaction with a range of *Ams* and *Tfr* genotypes predictably biased F_1 and F_2 family sex ratios (Table 3b).

The relative frequencies of the three male morphs, as well as local sex ratios, are known to influence male and female fitness in *P. sculpita*^{7,19,20}. Biases in male morph and sex ratios, moreover, arise and vanish without pattern within patchily distributed spongocoels^{6,20}. Conditional strategies are unlikely to evolve in such unpredictable environments^{21–24}. Thus, our model of interaction between *Ams*, *Tfr*, ECF and primary sex-determination loci is consistent with known aspects of this species' biology and with

established theory. Our results demonstrate the mendelian inheritance of male mating behaviour and sex factor loci, whose alleles interact with each other, as well as with an apparent extrachromosomal sex-ratio biasing factor. Interactions among these factors could rapidly shift population sex ratios in response to the dynamics of this species' mating system. Allele frequencies at *Ams* and *Tfr* conform to Hardy–Weinberg expectations⁷ (Table 3a), perhaps because fitness interactions between alleles at both loci cycle rapidly^{3,25}. Thus our results also demonstrate that genetic polymorphisms and epistasis affecting fitness can arise and persist in nature^{26,27}. □

Methods

Genetic crosses and electrophoresis. Male isopods and virgin females²⁸ were collected from sponges^{6,29} and maintained as pairs until females became gravid⁶. The F_1 generation included eight α -males, six β -males and five γ -males, each crossed to haphazardly selected females, to yield eight α -families, 13 β -families and 10 γ -families. We reared F_1 animals to maturity⁷ and produced an F_2 generation from six F_1 α -males, 13 F_1 β -males and four F_1 γ -males each crossed with F_1 daughters of β -males, as well as two F_1 β -males each crossed to F_1 daughters of γ -males. F_2 animals were reared under the same conditions as F_1 animals. Tissue samples from all adults of each generation were electrophoresed and stained for *Pgm* activity⁸.

Estimation of expected *Ams* frequencies for the F_1 . Previously described methods⁷ estimate that over 99% of field-collected individuals possess *Ams* ^{α} *Ams* ^{α} (0.86), *Ams* ^{β} *Ams* ^{α} (0.02) or *Ams* ^{γ} *Ams* ^{α} (0.11) genotypes⁷, limiting the possible allelic combinations for F_1 progeny. Females are evidently heterogametic in this and in related sphaeromatid species (S.M.S. and C.S., unpublished electrophoretic data)^{10,17,30}. Thus, we presumed that *P. sculpita* females carry alleles at the *Ams* locus at the same frequencies as described for males⁷, and we expected primary sex ratios to equal unity.

Detection of genetic interactions. The genetic model⁷ indicated that the *Pgm/Ams* complex and primary sex determination loci in *P. sculpita* are unlinked. Thus, a cross between a homogametic male, heterozygous at *Pgm/Ams*, and a heterogametic female, homozygous at *Pgm/Ams*, would yield four combinations of two sexes and two *Pgm* genotypes in equal frequency. Using P_1 to represent males and P_2 for females, Q_1 as *Pgm* allele class 1 and Q_2 as *Pgm* allele class 2, we plotted the four progeny genotypes in a 2×2 table (Fig. 1a) and examined deviations in the table using G -tests. We identified sex-ratio biases by comparing $\Sigma(P_1Q_i)$ with $\Sigma(P_2Q_i) = G_P$, deviations from mendelian expectations at *Pgm* by comparing $\Sigma(P_iQ_1)$ with $\Sigma(P_iQ_2) = G_Q$, and interactions between sex ratio and *Pgm* frequency by comparing $(P_1Q_1 + P_2Q_2)$ with $(P_1Q_2 + P_2Q_1) = G_{PQ}$. We predicted five possible patterns of G_P , G_Q and G_{PQ} deviations among the F_1 families (shown in Fig. 1): a, independent assortment between primary sex factors and the *Pgm/Ams* complex would yield four combinations of two sexes and two *Pgm* genotypes in equal frequency, no significant G_P , G_Q and G_{PQ} deviations, and indicate no effect of sex-ratio biasing factors; b, lethal factors causing mortality unrelated to sex would cause deficiencies in the frequencies of *Pgm* alleles, significant G_Q deviations, but no deviations in sex ratio (G_P), and no interaction between *Pgm* and sex (G_{PQ}); c, factors causing sex-limited mortality would cause consistent deficiencies in the frequency of one or the other sex, consistent G_P deviations, but no deviations in *Pgm* frequency (G_Q), and no interaction between *Pgm* and sex (G_{PQ}); d, factors causing one-way sex change (genetic males maturing as females, for example) would cause no deviation in *Pgm* frequencies (G_Q), but would generate consistent sex-ratio (G_P) deviations, as well as significant *Pgm*–sex interactions (G_{PQ}); e, factors causing two-way sex change would only show significant G_{PQ} interactions.

Effects of *Tfr* and ECF. We let *Tfr* be a diallelic, autosomal locus whose alleles (*Tfr*¹, *Tfr*²) interact with, but assort independently of, alleles at *Ams* and at primary sex-determination loci (Fig. 2). *Tfr*¹*Tfr*¹ was assumed to have no effect on males of any *Ams* genotype, and no effect on *Ams* ^{α} *Ams* ^{α} females. However, females bearing *Tfr*¹*Tfr*¹, as well as β - or γ -alleles at *Ams*, were assumed to mature as males, with phenotypes determined by their *Ams* allelic state. *Tfr*²*Tfr*² was assumed to have no effect on females of any *Ams* genotype, and no effect on β - or γ -males. However, *Ams* ^{α} *Ams* ^{α} males bearing *Tfr*²*Tfr*² were assumed to mature as females. *Tfr*¹*Tfr*² heterozygotes were assumed to affect

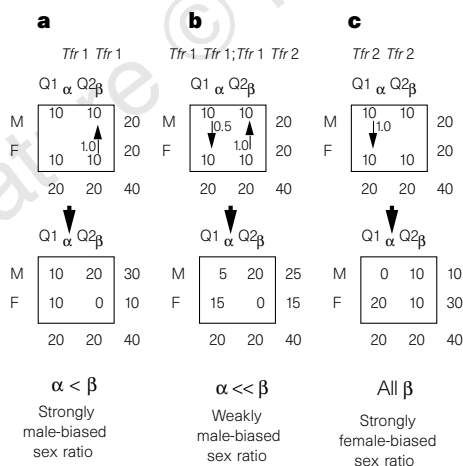


Figure 3 Effects of *Tfr* and ECF on mendelian inheritance at the *Ams* locus: upper 2×2 tables show hypothetical allelic combinations among 40 progeny from $Q_{1\alpha}Q_{1\alpha}$ female $\times$ $Q_{1\alpha}Q_{2\beta}$ male crosses (Q_{ij} , *Pgm/Ams* complex, where i is the *Pgm* allele and j , the *Ams* allele; M, males; F, females); small arrows indicate sex transformations resulting from epistatic and extrachromosomal interactions; *Tfr* genotypes are shown above each upper table; the numbers near the arrows indicate the proportions of transforming individuals; large arrows point to lower 2×2 tables, which show the results of sex transformations on α : β male morph and sex ratios; **a**, *Tfr*¹*Tfr*¹, $Q_{1\alpha}Q_{2\beta}$ females mature as males; **b**, *Tfr*¹*Tfr*², $Q_{1\alpha}Q_{1\alpha}$ males mature as females owing to ECF; all $Q_{1\alpha}Q_{2\beta}$ females mature as β males owing to *Tfr* and *Ams* interaction; **c**, *Tfr*²*Tfr*², $Q_{1\alpha}Q_{1\alpha}$ males mature as females.

only individuals descended from field-collected β -males, and then affect only two *Ams* genotypes: Tfr^1Tfr^2 females bearing β -alleles were assumed to mature as β -males, and Tfr^1Tfr^2 , $Ams^{\alpha}Ams^{\alpha}$ males were assumed to mature as females. This latter effect assumed that the Tfr^2 allele interacts with ECF, which initially occurred only in parental β -males, but which was transmitted to F_{1-2} individuals of both sexes and a range of *Ams* genotypes (Table 3b).

Testing the model. In Tables 3a,b, exact probabilities were Bonferroni-adjusted ($0.05/k$, where k is the number of tests) when multiple crosses with identical *Ams* and *Tfr* genotypes, as well as ECF states were tested; similar crosses with nonsignificant exact probabilities were pooled and the exact χ^2 probability for the pooled frequencies reported; primary sex-determination genotypes were unambiguously determined from *Pgm* genotype frequencies within families; the apparent *Tfr* genotypes among 36 parents (24 crosses) were 15 Tfr^1Tfr^1 , 12 Tfr^1Tfr^2 and 9 Tfr^2Tfr^2 ; expected genotypes calculated from inferred allele frequencies conform to Hardy–Weinberg expectations, exact χ^2 probability 0.21.

Received 21 January; accepted 24 April 1997.

- Lank, D. B., Smith, C. M., Hanotte, O., Burke, T. & Cooke, F. Genetic polymorphism for alternative mating behaviour in lekking male ruff. *Nature* **378**, 59–62 (1995).
- Gross, M. R. Alternative reproductive strategies and tactics; diversity within sexes. *Trends Ecol. Evol.* **11**, 92–97 (1996).
- Sinervo, B. & Lively, C. M. The rock–paper–scissors game and the evolution of alternative male strategies. *Nature* **380**, 240–243 (1996).
- Shuster, S. M. Alternative reproductive behaviors: Three discrete male morphs in *Paracerceis sculpita*, an intertidal isopod from the northern Gulf of California. *J. Crust. Biol.* **7**, 318–327 (1987).
- Shuster, S. M. The reproductive behaviour of α -, β -, and γ -males in *Paracerceis sculpita*, a marine isopod crustacean. *Behaviour* **121**, 231–258 (1992).
- Shuster, S. M. Male alternative reproductive behaviors in a marine isopod crustacean (*Paracerceis sculpita*): The use of genetic markers to measure differences in fertilization success among α -, β - and γ -males. *Evolution* **34**, 1683–1698 (1989).
- Shuster, S. M. & Wade, M. J. Equal mating success among male reproductive strategies in a marine isopod. *Nature* **350**, 606–61 (1991).
- Sassaman, C. Inbreeding and sex ratio variation in female-biased populations of a clam shrimp, *Eulimnadia texana*. *Bull. Mar. Sci.* **45**, 425–432 (1989).
- Hartl, R. & Clark, A. In *Principles of Population Genetics* 2nd edn, 47–57 (Sinauer, Sunderland, MA, 1989).
- Heath, D. J. & Ratford, J. R. The inheritance of sex ratio in the isopod, *Sphaeroma rugicauda*. *Heredity* **64**, 419–425 (1990).
- Bull, J. J. *Evolution of Sex Determining Mechanisms* (Benjamin-Cummings, Menlo Park, CA, 1983).
- Legrand, J. J., Legrand-Hamelin, E. & Juchault, P. Sex determination in Crustacea. *Biol. Rev.* **62**, 439–470 (1987).
- Juchault, P., Rigaud, T. & Mocquard, J.-P. Evolution of sex-determining mechanisms in a wild population of *Armadillidium vulgare* Latr. (Crustacea: Isopoda): competition between two feminizing parasitic sex factors. *Heredity* **69**, 382–390 (1992).
- Rousset, F., Bouchon, D., Pintureau, B., Juchault, P. & Solignac, M. *Wolbachia* endosymbionts responsible for various alterations of sexuality in Arthropods. *Proc. R. Soc. Lond. B* **250**, 91–98 (1991).
- Hurst, L. D. The incidences, mechanisms and evolution of cytoplasmic sex ratio distorters in animals. *Biol. Rev.* **68**, 121–193 (1993).
- Rigaud, T. & Juchault, P. Conflict between feminizing sex ratio distorters and an autosomal masculinizing gene in the terrestrial isopod, *Armadillidium vulgare* Latr. *Genetics* **133**, 247–252 (1993).
- Juchault, P. & Rigaud, T. Evidence for female heterogamety in two terrestrial crustaceans and the problem of sex chromosome evolution in isopods. *Heredity* **75**, 488–471 (1995).
- Read, T. R. C. & Cressie, N. A. C. In *Goodness-of-fit Statistics for Discrete Multivariate Data* 136–139 (Springer, New York, 1988).
- Shuster, S. M. & Wade, M. J. Female copying and sexual selection in a marine isopod crustacean. *Anim. Behav.* **42**, 1071–1078 (1991).
- Shuster, S. M. In *Crustacean Sexual Biology* (eds Bauer, R. T. & Martin, J. W.) 91–110 (Columbia Univ. Press, New York, 1991).
- Levins, R. *Evolution in Changing Environments* (Princeton Univ. Press, Princeton, 1968).
- Slatkin, M. On the equilibration of fitnesses by natural selection. *Am. Nat.* **112**, 845–859 (1978).
- Maynard Smith, J. *Evolution and the Theory of Games* (Cambridge Univ. Press, New York, 1982).
- Lively, C. M. Canalization versus developmental conversion in a spatially variable environment. *Am. Nat.* **128**, 561–572 (1986).
- Slatkin, M. The evolutionary response to frequency and density dependent interactions. *Am. Nat.* **114**, 384–398 (1979).
- Wright, S. *Evolution and Genetics of Populations* Vol. 2 (Univ. of Chicago Press 1969).
- Wade, M. J., Shuster, S. M. & Stevens, L. Bottlenecks, founder events and inbreeding: Experimental studies of the response to selection with *Tribolium*. *Evolution* **50**, 723–733 (1996).
- Shuster, S. M. Changes in female anatomy associated with the reproductive molt in *Paracerceis sculpita* (Holmes), a semelparous isopod crustacean. *J. Zool.* **225**, 365–379 (1991).
- Shuster, S. M. Courtship and female mate selection in a semelparous isopod crustacean (*Paracerceis sculpita*). *Anim. Behav.* **40**, 390–399 (1990).
- Tinturier-Hamelin, E. Sur le polychromatisme de l'isopode Flabellifere *Dynamene bidentata* (Adams) II. Etude genetique du mutant *bimaculata* partiellement. *Cah. Biol. Mar.* **4**, 473–591 (1967).

Acknowledgements. This research was supported by the NSF and by organized research and departmental funding from Northern Arizona University, and was authorized by the Mexican Government. We thank M. Wade and B. Charlesworth for reviewing data and earlier drafts of the manuscript; Y. Toquenaga for statistical advice and for a program for calculating exact χ^2 tests; D. Dorado, S. Juarez, S. Hag, H. Baitoo, S. Bhakta, Y. Bhakta, S. Brekhus; H. Yoon; N. Kim, M. Kim, U. Rao and L. Lynch for assistance in maintaining laboratory animals; and V. Jormalainen, P. Nelson, K. Johnson, G. Davis and M. Pitts for discussion.

Correspondence and requests for materials should be addressed to S.M.S. (e-mail stephen.shuster@nau.edu).

Different types of fear-conditioned behaviour mediated by separate nuclei within amygdala

Simon Killcross*, Trevor W. Robbins & Barry J. Everitt

Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK

The amygdala has long been thought to be involved in emotional behaviour^{1,2}, and its role in anxiety and conditioned fear has been highlighted^{3,4}. Individual amygdaloid nuclei have been shown to project to various cortical and subcortical regions implicated in affective processing^{5–7}. Here we show that some of these nuclei have separate roles in distinct mechanisms underlying conditioned fear responses. Rats with lesions of the central nucleus exhibited reduction in the suppression of behaviour elicited by a conditioned fear stimulus, but were simultaneously able to direct their actions to avoid further presentations of this aversive stimulus. In contrast, animals with lesions of the basolateral amygdala were unable to avoid the conditioned aversive stimulus by their choice behaviour, but exhibited normal conditioned suppression to this stimulus. This double dissociation demonstrates that distinct neural systems involving separate amygdaloid nuclei mediate different types of conditioned fear behaviour. We suggest that theories of amygdala function should take into account the roles of discrete amygdala subsystems in controlling different components of integrated emotional responses.

Investigations of the neural basis of pavlovian fear conditioning and its role in anxiety have suggested that the lateral nucleus of the amygdala is the site of convergence of neural pathways that carry information about conditioned stimuli (CSs) and aversive reinforcers (USs)^{4,6}. The emotional expression of this learned association may then be mediated by neural connections from the lateral to the central nucleus of the amygdala⁴, which, through its projections to hypothalamic and brainstem areas, is thereby able to coordinate the behavioural, endocrine and autonomic responses that form an integrated emotional response⁸.

Serial processing in these regions of the amygdala is known to be involved in the acquisition and expression of conditioned fear responses to aversive CSs, such as freezing and fear-potentiated startle in animals^{3–6}. However, the role of these regions in alternative indices of fear conditioning, including the instrumental choice responses involved in avoidance or conflict behaviour, is much less clear⁹. We therefore wished to investigate whether all forms of fear conditioning are mediated by this serial information flow between the lateral and basal nuclei to the central nucleus of the amygdala. To achieve this we designed a fear-conditioning procedure in rats in which the development of an aversive CS–US association could be assessed simultaneously by examination of two dissociable fear responses in the same animal. The aversive CS–US association created by this procedure would not only produce a pavlovian conditioned fear response, but would also provide the necessary information for animals to solve an operant discrimination that would lead to the avoidance of future presentations of the aversive stimulus.

Rats were trained on a concurrent conditioned-suppression and conditioned-punishment task (see Methods). Pressing one lever in an operant chamber produced an aversive conditioned stimulus (CS+), pressing the other lever produced control presentations of a

* Present address: Department of Psychology, University of York, Heslington, York YO1 5DD, UK.

neutral conditioned stimulus (CS–). In sham-operated control animals, the aversive CS+ caused a disruption of ongoing lever pressing during each of its presentations, relative to performance during the control CS– (Fig. 1, left). This conditioned-suppression effect is a well-established measure of the formation of a pavlovian aversive CS–US association. Behavioural evidence indicates that this is a pavlovian conditioned response deriving from species-specific defence responses¹⁰. Sham-operated control rats simultaneously came to bias their lever-press responses away from the lever producing the CS+ and towards the lever producing the neutral CS– (Fig. 1, right). This conditioned-punishment effect also demonstrates the presence of an intact aversive CS–US association, as it is only on the basis of this association that rats may choose between the levers and modulate their actions to avoid future aversive events¹¹. The concurrent assessment of conditioned suppression and conditioned punishment provided simultaneous, alternative measures of the establishment of the aversive CS–US association, the former dependent on a conditioned cessation of ongoing behaviour, and the latter on a choice between two available actions.

Lesions of the lateral and basal amygdala (Figs 2 and 3) did not affect the conditioned suppression of responses elicited by the CS+ (Fig. 1, left), but significantly impaired the ability of the animals to bias their choice of action away from the lever that produced the aversive, punishing stimulus (Fig. 1, right). In contrast, lesions of the central nucleus of the amygdala (Figs 2 and 3) produced severe impairments in the rats' suppression of baseline responses during presentations of the CS+ (Fig. 1, left), but their ability to direct their responses away from the lever that led to these presentations was unaffected (Fig. 1, right). Animals with lesions to both structures, as expected, showed a blockade of both conditioned-punishment and conditioned-suppression effects (Fig. 1).

This double dissociation of the effects of lesions of different amygdaloid nuclei on the behavioural expressions of fear conditioning has important implications for our understanding of the role of the amygdala in the formation of aversive CS–US associations. This is particularly important because the pattern of results cannot be explained by gross changes in primary motivation, such as sensitivity to footshock¹², or by deficits in stimulus discrimination (as both conditioned suppression and conditioned punishment are assessed using measures that contrast behaviour associated with the CS+ with that associated with the CS–). According to the hypothesis that the lateral nucleus of the amygdala is the site of formation of the CS–US association, which then achieves behavioural output through projections to the central nucleus of the amygdala^{4,13}, lesions of either the central nucleus or the basolateral amygdala should have produced complete and profound deficits in both conditioned suppression and conditioned punishment. Thus lesions of the basolateral amygdala should have prevented any formation of CS–US associations, whereas lesions of the central nucleus of the amygdala should have prevented such associations from being expressed in the behaviour. The double dissociation shown here demonstrates that both the basolateral amygdala and the central nucleus are not only independently capable of supporting the formation of CS–US associations, but also independently mediate the different forms of behavioural expression of defensive responses that are dependent on these associations.

Two strong conclusions follow from this dissociation. First, the lateral nucleus cannot be the only site at which pavlovian CS–US associations are stored, as rats with extensive lesions of the basolateral amygdala, which included the lateral nucleus, were shown to have intact CS–US associations as assessed by conditioned suppression. Although it has been shown that excitotoxic lesions of the basolateral amygdala produce deficits in the acquisition of conditioned fear responses, as assessed by freezing¹⁴, fear-potentiated startle¹³ and lick suppression¹², other findings have suggested that this deficit is abolished following more extended training¹⁵. Our rats with lesions in the basolateral amygdala also exhibited in the first

session of training a transitory deficit in suppression that was not apparent in any subsequent sessions, where the level of suppression did not differ from that of control animals. In contrast, animals with lesions of the central nucleus of the amygdala or combined lesions showed a persistent deficit across all sessions. Hence, although the basolateral amygdala may contribute to the formation of CS–US associations, the convergence of sensory information mediating conditioned fear can clearly also occur independently of this region, for example within the central nucleus of the amygdala or its afferent circuitry, including direct projections from midline thalamic nuclei and the sensory posterior thalamus¹⁶.

Second, our results are not compatible with theories suggesting that all manifestations of aversively motivated conditioned fear behaviour, including conflict and avoidance responses, are mediated by the various descending brainstem and hypothalamic projections of the central nucleus^{3,5,6}. This important conclusion follows because rats with complete lesions of the amygdaloid central nucleus, although impaired in pavlovian conditioned suppression,

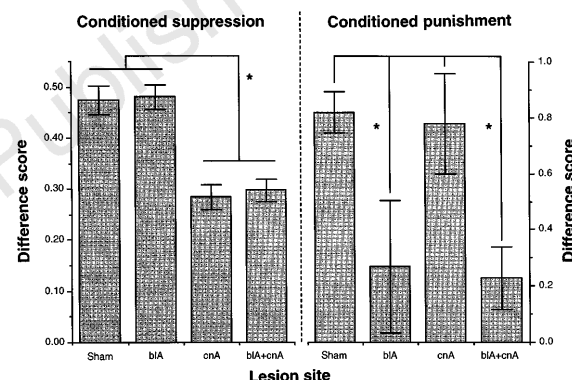


Figure 1 Left, Difference between suppression ratios during presentation of the aversive CS+ and neutral CS– in sham-operated control rats, and animals with lesions of the lateral/basolateral nuclei of the amygdala (bIA), the nucleus (cNA), or with combined lesions of both (cNA + bIA) averaged over 10 sessions. In no group did animals show significant suppression of lever pressing during CS– relative to periods without stimulus presentation. A difference score of 0.5 represents complete suppression during CS+ with no suppression during CS– (high fear conditioning); 0.0 represents no suppression during either CS+ or CS– (low fear conditioning). Rats with cNA or cNA + bIA lesions showed significantly reduced levels of fear conditioning relative to sham-operated controls and rats with lesions of just the bIA. Analysis of variance with factor 'lesion' (sham, bIA, cNA and bIA + cNA) indicated a significant effect of lesion ($F_{3,36} = 8.24, P < 0.001$). Pairwise comparisons showed that performance of cNA rats did not differ from that of rats with combined lesions, but both groups differed (asterisk, $P < 0.05$) from bIA rats and from sham-operated controls. Rats with lesions of the bIA did not differ from sham-operated controls. Right, difference in rate of lever pressing on the lever producing the neutral CS– and on the lever producing the aversive CS+. Lever pressing was measured when the aversive stimulus was not being presented, and is expressed relative to performance during baseline sessions in which no stimuli were presented. In no group did performance on the CS– lever differ significantly from baseline rates. A score of 1.0 represents a complete bias away from the CS+ lever while maintaining responses on the CS– lever at rates equivalent to baseline (high fear conditioning); 0.0 represents no difference between performance on the CS+ and CS– levers, with neither rate differing from baseline performance (low fear conditioning). Sham-operated controls and rats with cNA lesions showed a significant bias in responding away from the lever producing the aversive CS+; animals with bIA lesions or combined lesions showed no such bias. Analysis of variance with factor 'lesion' indicated a significant effect ($F_{3,36} = 3.42, P < 0.05$). Pairwise comparisons revealed that sham controls and rats with cNA lesions did not differ in lever discrimination (and hence degree of conditioned fear), but both showed significantly greater levels of conditioned fear (asterisk, $P < 0.05$) than animals with bIA lesions or combined lesions, which themselves did not differ.

showed normal formation of an aversive CS–US association as assessed by instrumental conditioned punishment behaviour. Thus the neural basis of any aversive CS–US association and its behavioural expression is distributed more widely in the brain. This finding is in line with previous data showing that, although important in freezing and startle behaviour, the central nucleus of the amygdala is unlikely to be important in active avoidance, conflict or punishment behaviours, or in aversive eyelid conditioning in the rabbit (which seems to depend on the deep nuclei of the

cerebellum and midbrain, including the interpositus and red nuclei¹⁷). Serial information transfer from the basolateral to the central nuclei of the amygdala seems to be critical only for certain classes of conditioned fear responses, assessed using specific behavioural output systems (indeed, the basolateral amygdala has been shown to be important in the unconditioned, as well as the conditioned, modulation of startle responding¹⁸), and does not seem to provide a general mechanism underlying fear conditioning. Similarly, just as different forms of conditioned fear responses seem

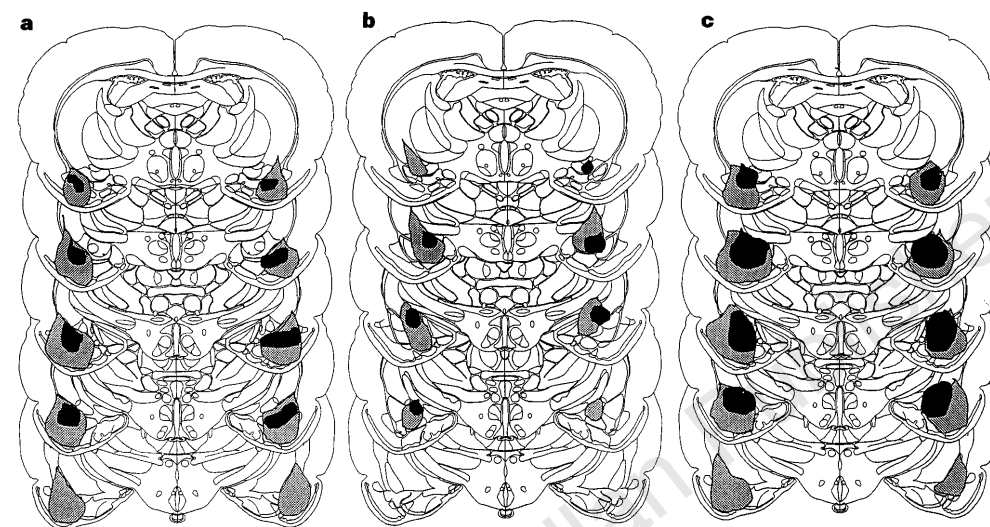


Figure 2 Representation of largest (pale shading) and smallest (dark shading) lesions of: **a**, lateral/basolateral nuclei; **b**, central nucleus; and **c**, both basolateral and central nuclei of the amygdala. Outlines are reproduced from ref. 30 and represent sections ranging from 1.33 to 3.9 mm posterior to Bregma.

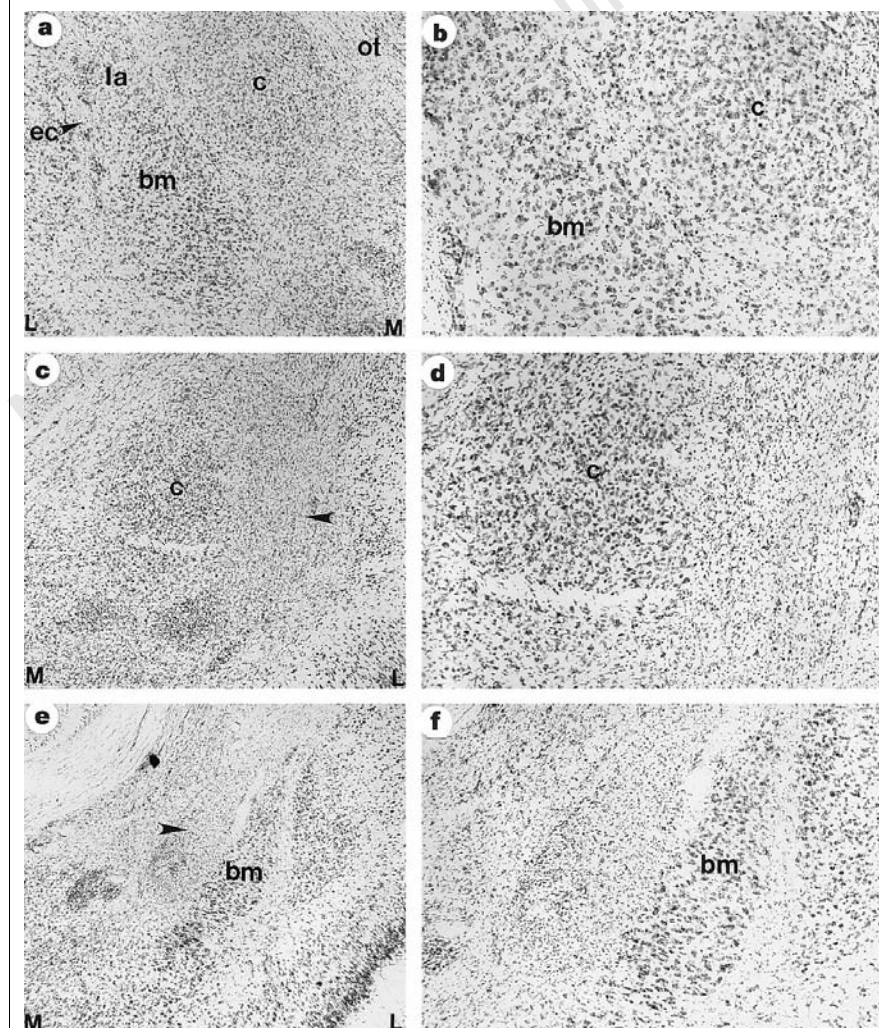


Figure 3 Photomicrographs of sections from three representative brains showing an intact brain (**a, b**) and excitotoxic lesions of the basolateral (**c, d**) and central (**e, f**) amygdala. **a, b**, A section (right side) through the amygdala of a sham-operated control at low (**a**) and higher (**b**) magnification. The appearance of the lateral (**la**) and basal magnocellular (**bm**) nuclei of the amygdala, which are together referred to as the basolateral amygdala, are visible, along with the central nucleus (**c**). The lateral and medial boundaries of the amygdala are formed by the external capsule (**ec**) and optic tract (**ot**), respectively. **c, d**, A section (left side) through the brain of a subject with a lesion of the basolateral amygdala at the same magnifications as **a**, and **b**. The distinctive magnocellular neurons of the basal magnocellular nucleus have been destroyed, so only glial nuclei are visible; the central nucleus has been spared (compare **c** with **a** and **d** with **b**). **e, f**, A section (left side) through the brain of a subject with a lesion of the central nucleus, at the same magnifications as **a** and **b**. The distinctive neurons of the central nucleus can no longer be seen; only glial nuclei remain stained in this area). The basolateral parts of the amygdala have been spared (compare **e** with **a** and **f** with **b**).

to depend on dissociable systems in the amygdala, assessments of pavlovian appetitive responding have failed to demonstrate a unitary role for amygdala nuclei in the development of simple appetitive CS-US associations. Rather, they implicate the basolateral amygdala in the control of behaviour by secondary reinforcers⁹, and they implicate the central nucleus of the amygdala in behaviour dependent on the predictive status of the CS itself, regardless of its association with primary reinforcement⁸.

Our results support the hypothesis³⁻⁷ that projections from the central nucleus of the amygdala are involved in conditioned responses elicited by aversive CSs. However, they also suggest that different projections from the basolateral amygdala are important in the production and direction of instrumental actions in response to fearful stimuli⁹. The demonstration that these dissociable systems can support different forms of fear-related behaviour, each dependent on the formation of CS-US associations, suggests that aversive responding in general may comprise aspects of behaviour derived from both systems. It is well established that projections from the central nucleus are important in reflexive emotional responding³⁻⁶. However, the role of direct projections from the basolateral amygdala to areas such as the ventral striatum and medial prefrontal cortex¹⁹ is not fully understood. These regions are important elements in the limbic cortico-ventral striatopallidal circuitry that provides an interface between the processing of emotionally salient stimuli and intentional action^{6,9}. For example, in humans, areas of the orbital prefrontal cortex that have significant reciprocal connections with the basolateral amygdala have been implicated in the assignment of affective or 'somatic' markers that inform choice behaviour^{20,21}. Moreover, our previous work on the neural mechanisms underlying the control of instrumental behaviour by appetitive CSs has demonstrated that interactions between the basolateral amygdala and the ventral striatum provide an important route by which associative processes gain access to instrumental response mechanisms^{9,22}. Our present results suggest that comparable interactions may underlie the effects of aversive conditioned punishers on aversive instrumental discrimination. Indeed, one way in which the present results may be interpreted is to regard the basolateral amygdala as part of a system responsible for voluntary or intentional instrumental choice behaviour based on emotional events, whereas the central nucleus of the amygdala is involved more closely in the reflexive, automatic pavlovian conditioned responses evoked by motivationally salient stimuli.

These results also have more direct implications for clinical research into the function of the amygdala in fear, anxiety and emotional behaviour in general. The amygdala in humans has been shown to be involved in varied aspects of fear conditioning²³, emotional memory^{24,25} and the recognition of emotion in facial expressions^{26,27} or vocal intonation²⁸. For example, our findings may be seen to provide neuroanatomical support for theories of anxiety that explain the heterogeneity of disorders by appealing to interactions between the dual mechanisms of reflexive pavlovian conditioned fear and voluntary avoidance behaviour, mediated by related neural systems²⁹. In the analysis of the effects of damage to the amygdala in humans²³, it may be important to consider the dissociable roles in different aspects of fear-related behaviour of nuclei of the amygdala and the subsystems of which they are part. □

Methods

Behavioural procedures. We trained 42 rats in operant chambers to maintain pressing on two levers for food on independent variable interval 60-s schedules. After lesions were made the rats were trained on a concurrent conditioned punishment and suppression task. Superimposed on the schedules of food reinforcement were two further independent variable interval 120-s schedules of a response-contingent 10-s auditory CS+ (3-kHz tone or 10-Hz clicks at 80 dB, counterbalanced) terminated by mild footshock (0.2 mA for 0.5 s) on one lever, and a matched neutral CS- (clicks or tone) on the other. Even with a

low response rate, the number of stimulus presentations on such schedules is effectively independent of response rate, so all animals received equivalent numbers of CS-US pairings and control CS presentations, regardless of their actual level of responding. In each of 10 30-min sessions the rats received an average of about 12 presentations of the CS+ and CS-. Rates of pressing on the two levers were recorded during baseline sessions in which no stimuli were presented, and in test sessions during the CS+ and CS- and in the absence of stimulus presentation, during the inter-trial interval.

Surgical procedures. The rats were anaesthetized using Avertin and received either bilateral excitotoxic lesions of the central nucleus of the amygdala with 0.063 M ibotenic acid (AP -2.2, -2.7, $L \pm 4.0$, $V -7.8$ (from Bregma), 0.2 μ l per infusion site), or bilateral excitotoxic lesions of the basolateral amygdala with 0.09 M quinolinic acid (AP -2.3, -3.0, $L \pm 4.6$, $V -7.3$ (from Bregma), 0.3 μ l per infusion site). Sham-lesioned animals received identical surgical treatment except for the infusion of excitotoxin. The choice of neurotoxins was based on extensive pilot experiments achieving discrete lesions of the two areas. Subsequent histological analysis revealed selective lesions of the central nucleus ($n = 6$), the lateral and basal magnocellular nuclei ($n = 9$), along with lesions of both structures ($n = 9$) and sham controls ($n = 18$).

Received 7 March; accepted 2 May 1997

- Weiskrantz, L. Behavioural changes associated with ablations of the amygdaloid complex in monkeys. *J. Comp. Physiol. Psychol.* **49**, 381-391 (1956).
- Kliver, H. & Bucy, P. C. Preliminary analysis of the temporal lobes in monkeys. *Arch. Neurol. Psychiat.* **42**, 979-1000 (1939).
- Davis, M. The role of the amygdala in fear and anxiety. *Annu. Rev. Neurosci.* **15**, 353-375 (1992).
- LeDoux, J. *The Emotional Brain* (Simon and Schuster, New York, 1996).
- Davis, M., Hitchcock, J. M. & Rosen, J. B. Anxiety and the amygdala: Pharmacological and anatomical analysis of the fear-potentiated startle paradigm. *Psychol. Learn. Motiv.* **21**, 263-305 (1987).
- LeDoux, J. E. Emotion: Clues from the brain. *Annu. Rev. Psychol.* **46**, 209-235 (1995).
- Kapp, B. S., Pascoe, J. P. & Bixler, M. A. in *The Neuropsychology of Memory* (eds Butters, N. & Squire, L. S.) 473-488 (Guildford, New York, 1984).
- Gallagher, M. & Chiba, A. A. The amygdala and emotion. *Curr. Opin. Neurobiol.* **6**, 221-227 (1996).
- Everitt, B. J. & Robbins, T. W. in *The Amygdala: Neurobiological Aspects of Emotion, Memory and Mental Dysfunction* (ed. Aggleton, J. P.) 401-429 (Wiley-Liss, New York, 1992).
- Bouton, M. E. & Bolles, R. C. Conditioned fear assessed by freezing and by the suppression of three different baselines. *Anim. Learn. Behav.* **8**, 429-434 (1980).
- Mowrer, O. H. & Solomon, L. N. Contiguity vs. drive-reduction in conditioned fear: The proximity and abruptness of drive reduction. *Am. J. Psychol.* **67**, 15-25 (1954).
- Selden, N. R. W., Everitt, B. J., Jarrard, L. E. & Robbins, T. W. Complementary roles for the amygdala and hippocampus in aversive conditioning to explicit and contextual cues. *Neuroscience* **42**, 335-350 (1991).
- Sananes, C. B. & Davis, M. N-Methyl-D-Aspartate lesions of the lateral and basolateral nuclei of the amygdala block fear-potentiated startle and shock-sensitization of startle. *Behav. Neurosci.* **106**, 72-80 (1992).
- Maren, S., Aharonov, G. & Fanselow, M. S. Retrograde abolition of conditional fear after excitotoxic lesions in the basolateral amygdala of rats: Absence of a temporal gradient. *Behav. Neurosci.* **110**, 718-726 (1996).
- Parent, M. B., Tomaz, C. & McGaugh, J. L. Increased training in an aversively motivated task attenuates the memory-impairing effects of posttraining N-Methyl-D-Aspartate-induced amygdala lesions. *Behav. Neurosci.* **106**, 789-797 (1992).
- Amaral, D. G., Price, J. L., Pitkanen, A. & Carmichael, S. T. in *The Amygdala: Neurobiological Aspects of Emotion, Memory and Mental Dysfunction* (ed. Aggleton, J. P.) 1-66 (Wiley-Liss, New York, 1992).
- Thompson, R. F. & Krupa, D. J. Organization of memory traces in the mammalian brain. *Annu. Rev. Neurosci.* **17**, 519-549 (1994).
- Wan, F. J. & Swerdlow, N. R. The basolateral amygdala regulates sensorimotor grating of acoustic startle in the rat. *Neuroscience* **76**, 715-724 (1996).
- Groenewegen, H. J., Berendse, H. W., Wolters, J. G. & Lohman, A. H. M. The anatomical relationship of the prefrontal cortex with the striatopallidal system, the thalamus and the amygdala: evidence for a parallel organization. *Prog. Brain Res.* **85**, 95-118 (1990).
- Damasio, A. *Descartes' Error* (Putnam, New York, 1994).
- Bechara, A., Tranel, D., Damasio, H. & Damasio, A. R. Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cereb. Cortex* **6**, 215-225 (1996).
- Everitt, B. J., Morris, K., O'Brien, A. & Robbins, T. W. The basolateral amygdala-central striatal system and conditioned place preference: Further evidence of limbic-striatal interactions underlying reward-related processes. *Neuroscience* **42**, 1-18 (1991).
- Bechara, A. et al. Double dissociation of conditioning and declarative knowledge relative to amygdala and hippocampus in humans. *Science* **269**, 1115-1118 (1995).
- Cahill, L., Babinsky, R., Markowitsch, H. J. & McGaugh, J. L. The amygdala and emotional memory. *Nature* **377**, 295-296 (1995).
- Cahill, L. et al. Amygdala activity at encoding correlated with long-term, free-recall of emotional information. *Proc. Natl Acad. Sci. USA* **93**, 8016-8021 (1996).
- Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372**, 669-672 (1994).
- Calder, A. J. et al. Facial emotion recognition after bilateral amygdala damage: Differentially severe impairment of fear. *Cogn. Neuropsychol.* **13**, 699-745 (1996).
- Scott, S. K. et al. Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature* **385**, 254-257 (1997).
- Mowrer, O. H. On the dual nature of learning—A re-interpretation of "conditioning" and "problem-solving". *Harvard Educ. Rev.* **17**, 102-148 (1947).
- Swanson, L. W. *Brain Maps: Structure of the rat brain* (Elsevier, Amsterdam, 1992).

Acknowledgements. We thank C. Morrison and H. Sweet for help with histology. This work was supported by a grant from the Human Frontiers Science Program, and a research fellowship to S.K. from Magdalene College, Cambridge.

Correspondence and requests for materials should be addressed to S.K. (e-mail: ask1@york.ac.uk).

Infants listen for more phonetic detail in speech perception than in word-learning tasks

Christine L. Stager & Janet F. Werker

Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia V6T 1Z4, Canada

Infants aged 4–6 months discriminate the fine phonetic differences that distinguish syllables in both their native and unfamiliar languages^{1–3}, but by 10–12 months their perceptual sensitivities are reorganized so that they discriminate only the phonetic variations that are used to distinguish meaning in their native language¹². It would seem, then, that infants apply their well honed phonetic sensitivities as they advance and begin to associate words with objects, but the question of how speech perception sensitivities are used in early word learning has not yet been answered. Here we use a recently developed technique to show that when they are required to pair words with objects, infants of 14 months fail to use the fine phonetic detail they detect in syllable discrimination tasks. In contrast, infants of 8 months—who are not yet readily learning words—successfully discriminate phonetic detail in the same task in which infants aged 14 months fail. Taken together, these results suggest a second reorganization in infants's use of phonetic detail as they move from listening to syllables to learning words.

To examine the phonetic information used by infants in early word learning, we used a recently developed procedure in which infants are first taught word–object pairings, and then tested on their ability to detect a change in either the word, the object, or both the word and object (J.F.W., L. B. Cohen, V. L. Lloyd, M. Casasola and C.L.S., manuscript in preparation). Using this procedure, it has been shown that when two-word–object pairings are used, infants aged 14 months, but not younger, are able to associate the dissimilar-sounding nonsense words with their referents (for example, infants learn object A is called “lif” and object B is called “neem”) (J.F.W. *et al.*, manuscript in preparation). This may be because younger infants can only easily learn and understand new words under conditions that provide rich contextual support^{4,5}.

To determine whether infants can use their speech-perception skills when first learning word–object associations, we tested 14-month-old infants in this procedure using phonetically similar nonsense words. If infants are using their fine phonetic discrimination skills, we would expect them to have no trouble learning two words that sound similar. In the first experiment, infants were presented with label A paired with object A and label B paired with object B (or vice versa, with order counterbalanced across infants). The two nonsense labels, “bih” and “dih”, were phonetically similar differing in only the place of articulation of the initial consonant. The two brightly coloured moving objects were created from modelling clay and matched for overall visual salience. The visual objects were presented on a monitor situated directly in front of the infant and the words were presented over a loudspeaker located below the monitor. During the habituation phase, the two word–object pairs were repeated in random order until the infant became familiarized to the pairings, as indicated by a criterial decline in looking time. Following habituation, two test trials were presented: one trial involved the same word–object combination (‘same’ trial) and the other involved the same words and objects, but a switch in the word–object pairing (‘switch’ trial) (Fig. 1).

We tested sixty-four 14-month-old infants (mean age, 14 months 16 days; range, 14 months 1 day to 14 months 29 days). Successful discrimination of the fine phonetic detail distinguishing these two

nonsense words would be revealed by infants noticing the switch in pairing and looking longer to the ‘switch’ trial than to the ‘same’ trial. A *t*-test revealed no significant difference in looking time [$t(63) = 0.916$; NS] ($M_{\text{same}} = 6.125$ s, $SD = 3.278$; $M_{\text{switch}} = 6.584$ s, $SD = 3.775$) (Fig. 2). Infants taught these minimally different words did not appear to notice the switch in word–object pairing.

In the second experiment, we investigated why infants fail to learn phonetically similar words. We tested infants on an even easier, single-word–object association task. We have shown that when the task is simplified to a single-word–object pairing, even much younger infants (8 months) can perform successfully (J.F.W. *et al.*, manuscript in preparation). This may be because the single-word–object association task is passable in two ways: infants who are readily able to link a word with an object may pass the task as a word-learning task, whereas infants who are not yet pairing the word and the object together can pass the task as a simple sound-discrimination task.

We tested sixteen 14-month-old infants (mean age, 14 months 12 days; range, 14 months 3 days to 14 months 30 days) and sixteen 8-month-old infants (mean age, 8 months 18 days; range, 8 months 4 days to 9 months 2 days). The same words and visual stimuli were used as in experiment 1, but infants were presented with only a single pair (for example, they were taught object A called “bih” and then tested with object A still called “bih” as the ‘same’ trial, and object A called “dih” as the ‘switch’ trial) (Fig. 1). When tested with these minimally different words even in this simplified procedure, infants of 14 months did not notice the switch in the word–object pairing. There was no significant difference in looking time to the ‘same’ trial as compared to the ‘switch’ trial [$t(15) = 0.242$; NS] ($M_{\text{same}} = 7.120$ s, $SD = 2.979$; $M_{\text{switch}} = 6.888$ s, $SD = 2.965$). However, infants of 8 months did notice the switch in word, and looked significantly longer during the ‘switch’ trial than during the ‘same’ trial [$t(15) = 2.225$, $P < 0.05$] ($M_{\text{same}} = 6.350$ s, $SD = 3.021$; $M_{\text{switch}} = 8.198$ s, $SD = 2.592$) (Fig. 2). These results indicate that the task is different for infants of 8 and 14 months of age: they suggest that for infants of 14 months, the single word–object association task does involve word learning, and that under these circumstances, the difference between “bih” and “dih” is not noticed. For infants of 8 months, this may be a simple sound-discrimination task, and the difference between “bih” and “dih” is easily detected.

To ensure that infants of 14 months are able to perform successfully

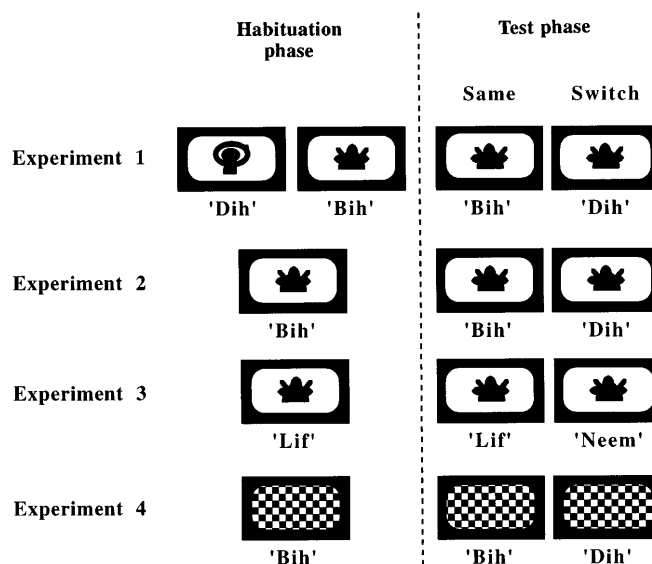


Figure 1 Diagrammatic representations of experiments 1–4.

in this single word-object association task, we did a control experiment using the phonetically dissimilar words that have previously been used in the two-word-object association task (J.F.W. *et al.*, in preparation) and using the visual objects we had chosen for this series of experiments. The procedure we used in experiment 3 was identical to that used in experiment 2, except that we used the phonetically dissimilar stimuli "lif" and "neem" as the object labels (Fig. 1). Sixteen 14-month-old infants (mean age, 14 months 6 days; range, 13 months 25 days to 14 months 29 days) completed this experiment. Infants clearly discriminated between the two labels [$t(15) = 5.412$, $P < 0.0001$] ($M_{\text{same}} = 7.984$ s, $SD = 2.769$; $M_{\text{switch}} = 12.428$ s, $SD = 2.294$) (Fig. 2). This result confirms that the difficulty experienced by 14-month-old infants in experiment 2 was due to the phonetic similarity of the labels used and not to other task factors.

The results from experiments 1–3 indicate that there is a difference in the phonetic detail used in word-learning versus speech-perception tasks, but the possibility also exists that infants of 14 months can simply no longer make the fine phonetic discriminations they demonstrated when younger. To eliminate this possibility, we did one further experiment with infants aged 14 months. We made the task one of simple speech discrimination by pairing the presentation of the words with a visual display that is unlikely to foster naming. We paired the word with the presentation of a checkerboard pattern on the TV monitor. There is considerable evidence that an unbounded stationary display like a checkerboard is unlikely to be perceived as an object by an infant⁶, and is thus less likely to be labelled⁷. All other aspects of the procedure were identical to those used in the previous experiment.

In this final study (experiment 4), sixteen 14-month-old infants (mean age, 14 months 20 days; range, 14 months 13 days to 14 months 28 days) were shown a checkerboard pattern paired with either the syllable "bih" or the syllable "dih" (Fig. 1). Under these testing conditions, infants of 14 months showed robust evidence of discriminating "bih" from "dih" [$t(15) = 3.691$, $P < 0.05$] ($M_{\text{same}} = 5.005$ s, $SD = 3.491$; $M_{\text{switch}} = 7.116$ s, $SD = 2.959$) (Fig. 2). This finding eliminates the possibility that infants of 14 months can no longer make fine phonetic discriminations. Instead, it presents compelling evidence that it is only when infants are attempting to learn the meaning of words that they fail to attend to the fine phonetic information.

Taken together, the results of these four experiments provide convincing evidence that infants use different information in word-learning than in speech-perception tasks. When listening for mean-

ing (experiments 1 and 2), infants of 14 months fail to detect the same phonetic detail that they can easily detect in a simple syllable discrimination task (experiment 4). Infants of 8 months may be able to pass the single-word-object association task used in experiment 2 because they are not mapping the sound onto meaning.

Inattention to phonetic detail may be beneficial to the child who is at the cusp of word learning. The task of linking words with objects is computationally more demanding than just listening to words as sounds. Thus, to be successful at this task, it may be necessary to limit the amount of detail, just as it is in other aspects of perceptual⁸ and linguistic⁹ development. At an older age, when word learning is no longer difficult, we would expect fine phonetic detail again to be accessed. By 3 years of age, English-learning children can distinguish most similar sounding words¹⁰. But at 14 months, limited phonetic information is all that is needed to avoid confusing a new word with the few words in the infants' lexicon¹¹.

The underlying pattern suggested by these results may be considered as another example of functional reorganization³. The decrease in the amount of detail used by infants as they move from speech perception to word learning is analogous to the earlier decline in infants' ability to discriminate non-native phones¹². In both cases, a decline rather than an increase in performance is evidence of a developmental progression. We believe that this type of functional reorganization is a necessary and ubiquitous aspect of development. Being aware of and understanding functional reorganizations may help us to resolve some of the puzzles of early development. □

Methods

The infant sat on the parent's lap facing the video screen. To prevent any biasing influence from seeing or hearing the display, parents wore a hat with a visor and a cloth blind and listened to female vocal music over headphones. The experimenter initiated trials when the infant visually fixated on a flashing red light in the centre of the monitor. The experimenter was seated in an adjoining control room, and recorded each infant's looking time on-line on a closed circuit TV system. The experimenter was unaware of the auditory stimulus being presented to the infant. Reliability coding conducted on 25% of the infants yielded reliability coefficients for each group of >0.90 . Across all four experiments, infants were excluded if they started to fuss during the experiment ($n = 32$), if they were not visible for on-line recording or filming ($n = 13$), if parents interfered during testing ($n = 9$), if the infant did not habituate ($n = 1$), if there was equipment failure ($n = 2$), or if infants did not look during one of the test trials ($n = 1$).

Received 16 January; accepted 2 May 1997.

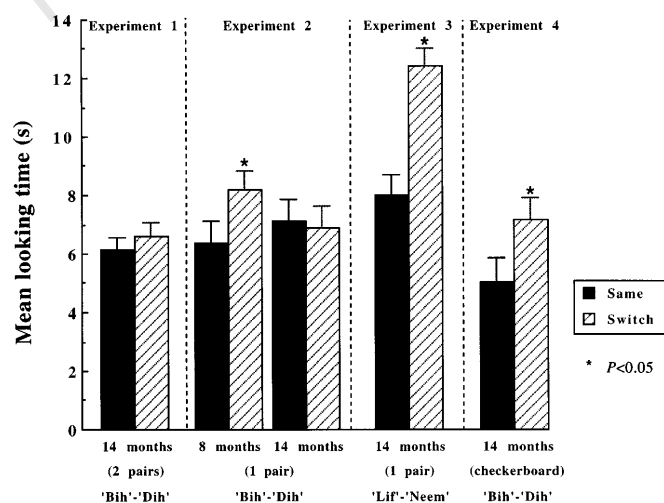


Figure 2 Results showing the conditions under which infants show significant recovery on the 'switch' trials. Graphs show mean looking times on the 'same' and 'switch' trials, with standard error bars.

- Kuhl, P. K. in *Handbook of Infant Speech Perception* Vol. 2 (eds Salapatek, P. & Cohen, L.) 275–382 (Academic, New York, 1987).
- Jusczyk, P. W. in *Handbook of Perception and Cognition, 11: Speech, Language, and Communication* (eds Miller, J. L. & Eimas, P. D.) 263–301 (Academic, Orlando, Florida, 1995).
- Werker, J. in *Invitation to Cognitive Science: Language* (eds Gleitman, L. R. & Liberman, M.) 87–106 (MIT Press, Cambridge, MA, 1995).
- Benedict, H. Early lexical development: Comprehension and production. *J. Child Lang.* **6**, 183–200 (1979).
- Oviatt, S. L. The emerging ability to comprehend language: An experimental approach. *Child Dev.* **51**, 97–106 (1980).
- Spelke, E. S., Vishton, P. & Van Hofsten, C. in *The Cognitive Neurosciences* (ed. Gazzaniga, M.) 165–179 (MIT Press, Cambridge, MA, 1994).
- Woodward, A. in *Twenty-fourth Annual Child Language Research Forum* (ed. Clark, E.) 35–45 (Cambridge University Press, Cambridge, MA, 1993).
- Turkewitz, G. & Kenny, P. A. Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. *Dev. Psychobiol.* **15**, 357–368 (1982).
- Newport, E. Maturation constraints on language learning. *Cog. Sci.* **14**, 11–28 (1990).
- Barton, D. *The Development of Communication* (eds Waterson, N. & Snow, C.) 255–261 (Wiley, New York, 1978).
- Charles-Luce, J. & Luce, P. A. An examination of similarity neighborhoods in young children's receptive vocabularies. *J. Child Lang.* **22**, 727–735 (1995).
- Werker, J. F. & Tees, R. C. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.* **7**, 49–63 (1984).

Acknowledgements. We thank J. A. R. Carr and R. C. Tees for comments on earlier drafts, and M. Bhatnager for her help in testing infants. We also thank the parents and infants who participated in the study. This research was supported by a Natural Sciences and Engineering Research Council of Canada Grant to J.F.W. and a NSERC Doctoral Scholarship to C.L.S.

Correspondence and requests for materials should be addressed to C.L.S. (e-mail: cstager@unixg.ubc.ca) or J.F.W. (e-mail: jwerker@cortex.psych.ubc.ca).

Laminar fine structure of frequency organization in auditory midbrain

Christoph E. Schreiner* & Gerald Langner†

* W. M. Keck Center for Integrative Neuroscience, Sloan Center for Theoretical Neuroscience, Coleman Laboratory, 513 Parnassus Avenue, University of California at San Francisco, San Francisco, California 94143-0732, USA

† Zoologisches Institut, Technische Hochschule Darmstadt, D-64287 Darmstadt, Germany

The perception of sound is based on signal processing by a bank of frequency-selective auditory filters, the so-called critical bands^{1–3}. Here we investigate how the internal frequency organization of the main auditory midbrain station, the central nucleus of the inferior colliculus (ICC), might contribute to the generation of the critical-band behaviour of its neurons. We find a unique spatial arrangement of the frequency distribution in the ICC that correlates with psychophysical critical-band characteristics. Systematic frequency discontinuities along the main tonotopic axis, in combination with a smooth frequency gradient orthogonal to the main tonotopic organization of cat ICC, reflect a layering of the frequency organization paralleling its anatomical laminae. This layered frequency organization is characterized by constant frequency ratios of corresponding locations on neighbouring laminae and may provide a spatial framework for the generation of critical bands and for signal processing within⁴ and across⁵ frequency bands for the analysis of sound.

Various perceptual phenomena indicate that the auditory system processes its inputs in the frequency domain by frequency-limited 'critical bands'^{1,2}. The importance of a frequency-based sound analysis is reflected in the tonotopic layout of the majority of auditory nuclei. Although the concept of critical bands is of fundamental importance in psychophysics, the mechanisms underlying the generation of the corresponding physiological band-pass filters are unknown. These frequency bands represent narrow spectral ranges within which sound energy is integrated and are crucial for a sound's perceived loudness³, its detectability in background noise⁴, and its discriminability from other sounds. For example, two simultaneously presented pure tones may only be heard separately when located in different critical bands.

A neuronal theory of critical bands would have to account for two essential and invariant filter attributes: a logarithmic relationship between the filter's bandwidth and centre frequency, and a level-independent bandwidth^{1,2}. It has been assumed that the filter processes of the auditory periphery are related to the critical-band concept³, but as the filter bandwidths of the cochlea and peripheral neurons are level-dependent and vary over a wide range^{6–14}, the required invariances of the critical-band filter can only be derived more centrally through processing mechanisms such as lateral inhibition. This mechanism creates level-tolerant filters in the central nervous system (CNS) of echolocating bats and has been encountered throughout the central auditory system^{15,16}. The lowest auditory station where neurons do have invariant filter properties comparable to critical bands is the ICC^{6–13}.

What features of ICC organization might contribute to the realization of critical band filters? The best-defined functional feature of organization of the ICC is the tonotopic representation of cochlear place or stimulus frequency^{17,18} along the dorsal–ventral extent of the ICC. Frequency mapping studies in the three dimensions of the ICC have led to the concept of functional 'iso-frequency laminae' orthogonal to the main tonotopic axis: that is, continuous sheets of neurons tuned to the same frequency that span the entire

ICC from edge to edge^{17–19}. Anatomical studies indicate that a prominent structural feature of the ICC is a preferred orientation of its principal neurons which are arranged in distinctive, parallel fibrodendritic laminae with a thickness of 120 to 180 μm (refs 20–25). In the cat ICC, this allows for 30 to 45 stacked anatomical laminae. They seem to be of functional significance because laminae defined by electrophysiology^{17–19,25,26}, by activity-dependent markers^{27,28}, or by neuroanatomical methods^{20–25} are of similar spatial orientation and extent.

We consider that a layered but largely continuous frequency organization exists in the ICC and that it is the neuronal substrate at which critical bands emerge. The spatial relationship of a laminated anatomical structure and a specific distribution of functional properties within and across laminae, such as complex local inhibition and integration network, appears to be particularly suited for the generation and function of critical bands. If lamination helps to create critical bands, three conditions must be fulfilled: (1) frequency relationships between neuronal neighbours on adjacent laminae have to be correlated with the size of critical bands; (2) short-range inhibition between adjacent laminae has to

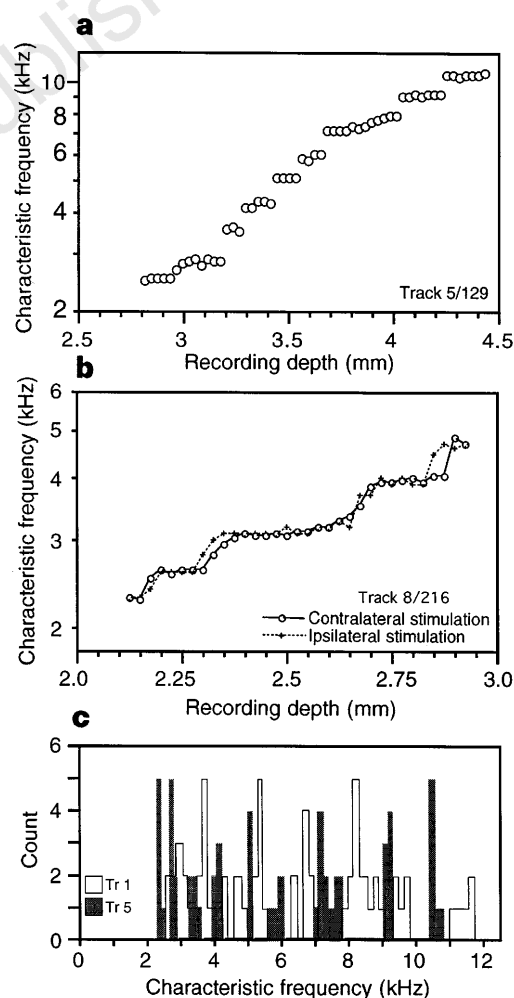


Figure 1 Stepwise progression of characteristic frequency (CF) along the main tonotopic axis in ICC. **a**, Example of a single electrode track with pronounced CF plateaux and steps. Note the similar step size on the logarithmic scale over 2 octaves. **b**, Example of CFs for contra- and ipsilateral stimulation. CF values obtained by stimulation of either ear are largely the same. Steps for the CF curves obtained for one ear may occur at slightly different depths from those for the other ear. **c**, Histogram of CF counts gathered from two parallel electrode tracks separated mediolaterally by $\sim 425 \mu\text{m}$. The histograms reveal a complementary frequency distribution by showing the same step sizes in the two penetrations but different frequency plateaux.

Figure 2 Spatial and frequency spacing of frequency plateaux along dorso-ventral electrode tracks in the ICC. **a**, Distribution of the change in recording depth from one frequency plateau to the next. The mean step size corresponds roughly to the spacing of anatomical laminae in the ICC²⁰⁻²⁵ (bar above the distribution). Only the spacing of consecutive frequency plateaux with at least three constant CF values each were included. **b**, Distribution of frequency step sizes in octaves. The range of behaviourally determined critical bandwidths^{8,9,14} (bar above the distribution) corresponds to the range of frequency steps. **c**, Frequency-dependence of step size. Circles represent frequency step size as a function of CF ($n = 8$ animals). Behaviourally determined critical bandwidths (CB) (open triangles⁸, filled triangles¹⁴) reveal the same trend and correspond to 1–2 frequency steps. Neuronal critical bandwidth (solid line represents average of bandwidth determined for single units with narrow-band noise masker⁷) matches the frequency dependency of the collicular frequency steps, as well as the behavioural data. The lower boundary of these neuronal data is similar to the behavioural as well as the frequency-step estimates of critical bands^{7,10}.

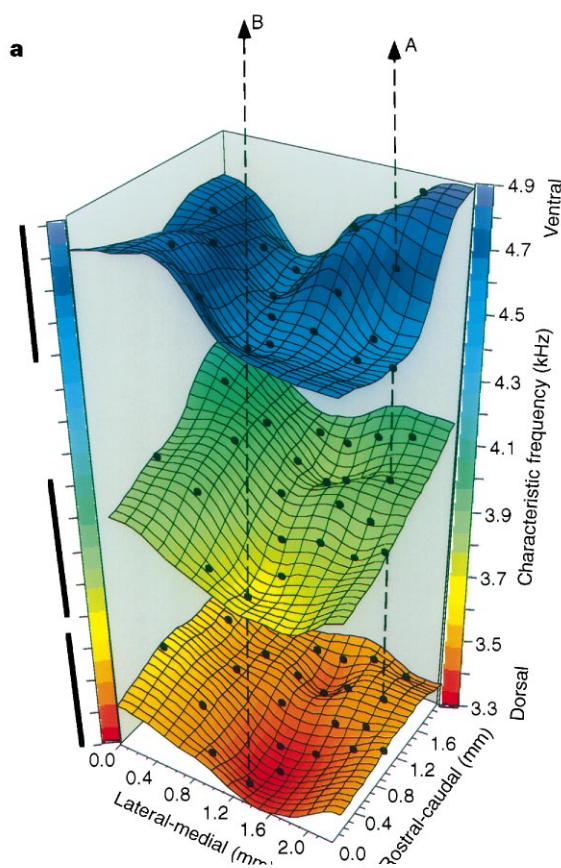
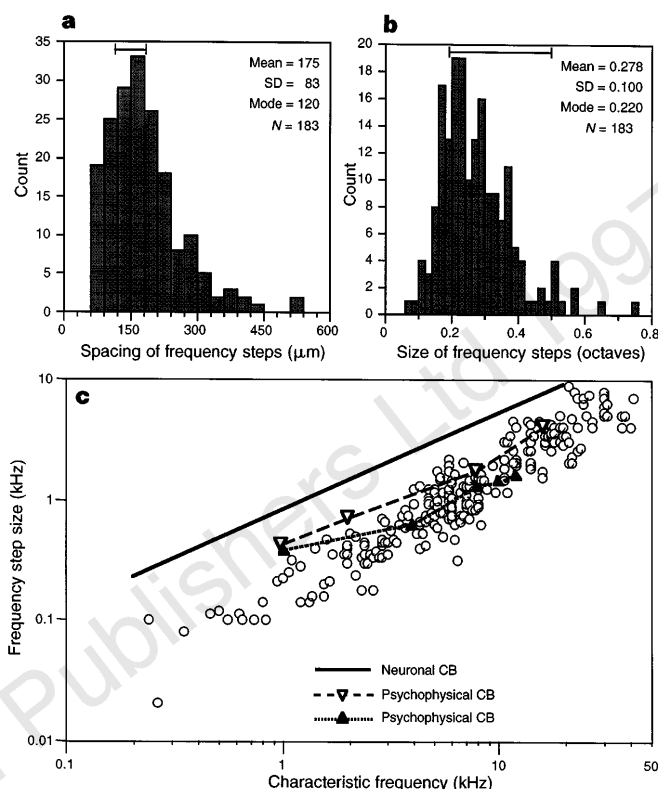
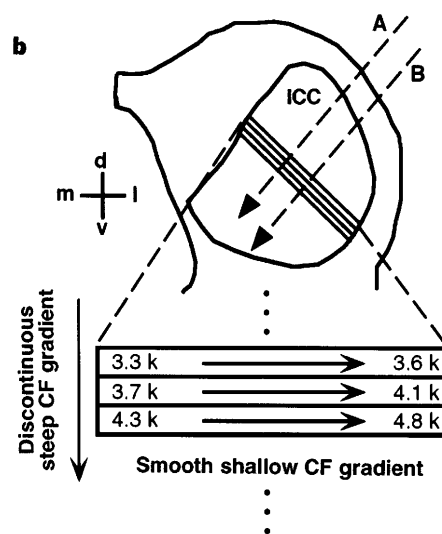


Figure 3 a, Reconstruction of functional frequency-band lamina in the ICC. Frequency representation in three adjacent frequency-band laminae was reconstructed from 23 parallel dorso-ventral electrode tracks in the ICC. The trajectory of two electrode tracks (A,B) is indicated by the arrows. The frequency content of each functional lamina was assigned by closest neighbourhood relation. Each lamina covers about 1/4 of an octave of non-overlapping frequencies (as indicated by the colour code and the bars on the left margin). In addition, for each lamina a small, approximately rostral-medial to caudal-lateral gradient of the tonotopic fine structure can be discerned with the aid of a pseudo-3D grid. Almost the full frequency range covered (3.3–4.9 kHz) is represented as the CF value on one of the 3 frequency-band laminae. The frequency gradient is not complete along the main dorsal-ventral tonotopic axis of the nucleus, but is



expressed as a layered sheet with a shallow gradient within each lamina and a steep gradient across laminae. The frequency gaps between the sampled frequency-band laminae are due to the fact the 23 tracks covered only ~4 mm² or ~2/3 of the total area of the full laminae. Given a systematic frequency representation within the laminae, not all expected CFs could be encountered in this incomplete mapping. **b**, Frequency gradients in the ICC. Arrows (A,B) indicate the trajectories of the corresponding example penetrations in **a**. Each schematic lamina displays the CFs at its lateral (l) and medial (m) edges for a fixed rostral-caudal location (0.4 mm). The full CF range within each lamina is slightly larger than indicated owing to an additional gradient in the rostral-caudal dimensions, and is ~3.1–3.7 kHz, 3.7–4.3 kHz and 4.3–5.2 kHz, respectively. d, Dorsal; v, ventral.

exist; and (3) the first two conditions have to be realized for each potential centre frequency of a critical band, that is, in principle, for all frequencies of the hearing range.

The second condition is probably fulfilled: inhibition between neurons in the ICC has been investigated in some detail and the existence of inhibitory connections between adjacent laminae has been demonstrated in bats and cats^{29,30}, although the precise role of inhibition on spectral receptive fields remains unclear. Anecdotal observation of discontinuities in the frequency representation^{17,18} suggests that the functional organization within the ICC might be related to its structural lamination^{25,27}. To reveal the potential relationship between laminae, frequency organization, and critical bands, a detailed analysis of the fine structure of the frequency organization in the dimension of the 'iso-frequency laminae' as well as along the main tonotopic axis is necessary.

The characteristic frequency (CF) of the locally most sensitive collicular neurons was obtained every 25–30 μm along latero–dorsal to medio–ventral microelectrode penetrations through the ICC of the cat. The estimated group characteristic frequencies for stimulation of either ear showed a global increase with recording depth in the latero–dorsal to medio–ventral direction, consistent with the previously defined tonotopic organization of the ICC^{17,18}. However, the progression of the characteristic frequency was not gradual, but discontinuous or stepwise^{17,18} over a large range of each penetration (Fig. 1). An analysis of the discontinuities revealed that the stepwise progression in the characteristic frequency sequence was characterized by two attributes: the spatial distance over which nearly constant characteristic frequencies were encountered, and the size of the frequency steps from one frequency 'plateau' to the next. The distance over which a given group of neurons dominated the characteristic frequency was $175 \pm 83 \mu\text{m}$ ($n = 183$) for penetrations that were roughly orthogonal to the orientation of the laminae (Fig. 2a). As the orientation of the laminae changes slightly with recording depth, the average distance measured in straight electrode tracks is larger than the actual distance between laminae. The relative stepwise increase of characteristic frequency was nearly invariant and, on average, 0.278 ± 0.100 octaves per frequency step (Fig. 2b) for the full audible frequency range. At CF = 4 kHz, for example, this corresponds to a step size of 0.8 kHz. Consequently, values of the characteristic frequency are not smoothly and completely distributed in penetrations along the main dorso–ventral tonotopic axis of the ICC but show a regularly spaced, discrete frequency representation.

Parallel electrode penetrations that were made close to each other ($<300 \mu\text{m}$) usually showed similar step sizes and plateau frequencies. In more widely spaced penetrations, the relative step sizes were also similar, but the plateau frequencies of the discrete frequency progression often differed by a small, constant amount. Accordingly, the frequencies encountered in more widely separated penetrations were systematically interspersed (Fig. 1c). It follows that the characteristic frequencies represented along a single dorso–ventral trajectory across the ICC do not continuously cover the complete frequency range. However, a non-quantal, complete characteristic frequency distribution can be established by combining all available dorso–ventral trajectories. The similar spacing of frequency discontinuities and fibro–dendritic laminae^{20–25} strongly suggests a correspondence between functional and structural laminar organization.

Additional support for a fine structure of the tonotopic organization in the ICC comes from a comparison of the characteristic frequency sequences for contra- and ipsilateral acoustic stimulation (Fig. 1b). Although the encountered characteristic frequencies along a given penetration generally matched for input from either ear, the depth location of the frequency discontinuities for the two ears were occasionally shifted by 30 to 90 μm . This suggests that binaural inputs of the same frequency are not always in perfect spatial register and provides physiological support for the anatomi-

cal finding that binaural inputs to the ICC may be spatially segregated^{20–25}.

These results confirm the hypothesis that the characteristic frequencies within a given lamina are not constant but change with location, and that the notion of 'frequency-band laminae' in the ICC is more appropriate than 'iso-frequency laminae'^{17–19}. More important, the findings show that neighbouring locations on adjacent laminae have a specific characteristic frequency ratio. The mode of these ratios was 0.22 octaves (for example, 4.66/4.0 kHz) which corresponds to ~ 0.5 –1 critical band in cats^{4,8,9,14} (Fig. 2c). Owing to the recording method, this finding only holds for low-threshold neurons and other groups of neurons may show deviations from this pattern^{19,21}.

The three-dimensional reconstruction of individual 'frequency-band laminae' revealed a systematic and continuous representation of a narrow frequency range within each lamina, manifesting as a gradient orthogonal to the main dorso–ventral tonotopic axis. The shallow gradient of this frequency representation is $\sim 0.09 \pm 0.04$ octaves mm^{-1} (Fig. 3a). The higher frequencies in the reconstructed laminae were generally located near the lateral edge of the nucleus and the lower frequencies near the rostro–medial edge (Fig. 3a,b). The frequency range across the extent of a lamina corresponds to the size of the frequency step to the next lamina.

These findings suggest the existence of a unique and precisely organized fine-structure of the frequency distribution in the ICC that consists of a segmented and layered characteristic frequency arrangement. In a stack of 30–45 frequency-band laminae (Fig. 3b), each exhibits a shallow, continuous frequency progression orthogonal to the traditionally defined main dorso–ventral tonotopic axis, and a more than ten-times steeper, but discontinuous frequency progression across the laminae ($\sim 1.6 \pm 0.7$ octaves mm^{-1} along the main tonotopic axis). We interpret this layered frequency organization as a potential structural substrate for the creation of critical bands by lateral inhibition. The ratios of the centre frequency of a neuron and of its neighbours on the two adjacent laminae is constant for all neurons, allowing reciprocal inhibitory interactions to shape the upper and lower boundaries of tuning curves appropriate for critical-band filters. The spatial relationships in the laminated frequency organization provide the necessarily constant relative bandwidths across the full range of audible frequencies. Accordingly, the three conditions deemed necessary for the generation of critical bands (see above) appear to be fulfilled by the layered fine structure of the tonotopic map in the ICC. Along the rostro–caudal extent of a frequency-band lamina, isofrequency contours (connecting locations of neurons with the same characteristic frequency) allow for the spatial representations of additional functional parameters. Consequently, the laminar organization of collicular neurons may be essential not only for the generation of critical bands but also for the extraction of information-bearing parameters for the perception of periodicity pitch and sound localization, as well as other perceptual attributes^{23,25,26}. □

Methods

The inferior colliculus (ICC) was exposed in 8 barbiturate-anaesthetized cats through a dorsal–lateral approach. Microelectrode penetrations were tilted laterally at a constant angle between 20° and 45° from vertical to achieve trajectories nearly orthogonal to the orientation of the anatomical laminae²¹. Multiple-unit responses were recorded every 25–30 μm and frequency response areas (FRAs) were reconstructed from 675 different level/frequency combinations. The characteristic frequency (CF) of the local group of neurons was defined as that frequency that produces the strongest neuronal response 5 to 10 dB above the lowest sound pressure producing any response. Accordingly, the CF data are biased towards the most sensitive local neurons. Often there were two response-threshold minima corresponding to two dominant CFs in the sampled cell population. With small advancement of the electrode, the CFs of the minima did not significantly change, whereas their threshold changed disproportionately with recording depth: one CF emerged at an increasingly

lower threshold and the other CF had an increasingly higher threshold. The occurrence at regular depth intervals of transitions from single-peaked to double-peaked tuning curves can be interpreted as a characteristic of functional lamination: single-peaked FRAs are representative of a spectrally uniform group of neurons within a functional lamina; double-peaked FRAs correspond to locations recording from neurons of two adjacent laminae. The fine spatial resolution necessary for defining a relationship between laminar and frequency organization, and the reconstruction of the frequency content of individual laminae, made it necessary to record activity from groups of neurons because it is not feasible consistently to isolate single neurons with the required spatial resolution. The same CFs and frequency steps were seen for advancing and retracting the electrode and for different electrode types.

A given 'frequency-band lamina' was reconstructed from the depth profile of several recording sites in three steps: (1) a CF plateau on a central penetration was chosen as the seed frequency for a functional lamina; (2) the horizontally closest penetration was selected and its CF plateau nearest to the seed frequency was determined, representing the most probable member of the same functional lamina as the seed frequency; (3) step 2 is repeated for the next nearest penetration until all penetrations have been included. Alignment for one lamina automatically defines the frequency content for all neighbouring laminae. In principle, the result of the alignment process does not depend on the initial seed frequency or penetration location.

Received 10 December 1996; accepted 2 May 1997.

1. Fletcher, H. Auditory patterns. *Rev. Mod. Phys.* **12**, 47–65 (1940).
2. Zwicker, E. et al. Critical bandwidth in loudness summation. *J. Acoust. Soc. Am.* **29**, 548–557 (1957).
3. Greenwood, D. D. Critical bandwidth and the frequency coordinates of the basilar membrane. *J. Acoust. Soc. Am.* **33**, 1344–1356 (1961).
4. Watson, C. S. Masking of tones by noise for the cat. *J. Acoust. Soc. Am.* **62**, 167–172 (1963).
5. Yost, W. A. & Sheft, S. Across-critical band processing of amplitude modulated tones. *J. Acoust. Soc. Am.* **85**, 848–857 (1989).
6. Ehret, G. & Merzenich, M. M. Auditory midbrain responses parallel spectral integration phenomena. *Science* **227**, 1245–1247 (1985).
7. Ehret, G. & Merzenich, M. M. Complex sound analysis (frequency resolution filtering and spectral integration) by single units of the inferior colliculus of the cat. *Brain Res.* **13**, 139–163 (1988).
8. Pickles, J. O. Normal critical bands in the cat. *Acta Otolaryngol.* **80**, 245–254 (1975).
9. Pickles, J. O. & Comis, S. D. Auditory-nerve-fiber bandwidths and critical bandwidths in the cat. *J. Acoust. Soc. Am.* **60**, 1151–1156 (1976).
10. Ehret, G. Critical bands and filter characteristics in the ear of the house mouse (*Mus musculus*). *Biol. Cybern.* **24**, 35–42 (1976).
11. Ehret, G. & Moffat, A. J. M. Noise masking of tone responses and critical ratios of the mouse cochlear nerve and cochlear nucleus. *Hearing Res.* **14**, 45–57 (1984).
12. Pickles, J. O. Psychophysical frequency resolution in the cat as determined by simultaneous masking and its relation to auditory-nerve resolution. *J. Acoust. Soc. Am.* **66**, 1725–1732 (1979).
13. Evans, E. F. et al. Correspondence between behavioural and physiological frequency selectivity in the guinea pig. *Br. J. Audiol.* **23**, 151–152 (1989).
14. Nienhuys, G. W. & Clark, G. M. Critical bands following the selective destruction of cochlear inner and outer haircells. *Acta Otolaryngol.* **88**, 350–358 (1979).
15. Suga, N. & Tsuzuki, K. Inhibition and level-tolerant frequency tuning in the auditory cortex of the mustached bat. *J. Neurophysiol.* **53**, 1109–1145 (1985).
16. Suga, N. Sharpening of frequency tuning by inhibition in the central auditory system: tribute to Yasuji Katsuki. *Neurosci. Res.* **21**, 287–299 (1995).
17. Merzenich, M. M. & Reid, M. D. Representation of the cochlea within the inferior colliculus of the cat. *Brain Res.* **77**, 397–415 (1974).
18. Roth, G. L. et al. Some features of the spatial organization of the central nucleus of the inferior colliculus of the cat. *J. Comp. Neurol.* **182**, 661–680 (1978).
19. Semple, M. N. & Aitkin, L. M. Representation of sound frequency and laterality by units in the central nucleus of the cat inferior colliculus. *J. Neurophysiol.* **42**, 1626–1639 (1979).
20. Rockel, A. J. & Jones, E. G. The neuronal organization of the inferior colliculus of the adult cat: I. The central nucleus. *J. Comp. Neurol.* **147**, 11–60 (1973).
21. Oliver, D. L. & Morest, D. K. The central nucleus of the inferior colliculus in the cat. *J. Comp. Neurol.* **222**, 237–264 (1984).
22. Oliver, D. L. Dorsal cochlear nucleus projections to the inferior colliculus in the cat: A light and electron microscope study. *J. Comp. Neurol.* **224**, 155–172 (1984).
23. Shneiderman, A. & Henkel, C. K. Banding of lateral superior olivary nucleus afferents in the inferior colliculus: a possible substrate for sensory integration. *J. Comp. Neurol.* **266**, 519–534 (1987).
24. Malmierca, M. S. et al. The central nucleus of the inferior colliculus in rat: A Golgi and computer reconstruction study of neuronal and laminar structure. *J. Comp. Neurol.* **333**, 1–27 (1993).
25. Chan, J. L. & Yin, T. C. T. Topographical relationships along the isofrequency laminae of the cat inferior colliculus: correlation with the anatomical lamination and representation of binaural response properties. *Soc. Neurosci. Abstr.* **8**, 348 (1982).
26. Schreiner, C. E. & Langner, G. Periodicity coding in the inferior colliculus of the cat. II. Topographical organization. *J. Neurophysiol.* **60**, 1823–1840 (1988).
27. Serviere, J. & Webster, W. R. A. A combined electrophysiological and [¹⁴C]-2-deoxyglucose study of the frequency organization of the inferior colliculus of the cat. *Neurosci. Lett.* **27**, 113–118 (1981).
28. Ehret, G. & Fischer, R. Neuronal activity and tonotopy in the auditory system visualized by c-fos gene expression. *Brain Res.* **567**, 350–354 (1991).
29. Pollak, G. D. & Park, T. J. The effects of GABAergic inhibition on monaural response properties of neurons in the mustache bat's inferior colliculus. *Hearing Res.* **65**, 99–117 (1993).
30. Palombi, P. S. & Caspari, D. M. GABA inputs control discharge rate primarily within frequency receptive fields of inferior colliculus neurons. *J. Neurophysiol.* **75**, 2211–2219 (1996).

Acknowledgements. Supported by the Office of Naval Research (C.E.S.) and the Deutsche Forschungsgemeinschaft (G.L.). We thank J. A. Winer for comments on the manuscript.

Correspondence and requests for materials should be addressed to C.E.S. (e-mail: chris@phy.ucsf.edu).

Kinesin hydrolyses one ATP per 8-nm step

Mark J. Schnitzer*† & Steven M. Block†‡

Departments of * Physics and † Molecular Biology, and ‡ Princeton Materials Institute, Princeton University, Princeton, New Jersey 08544, USA

Kinesin is a two-headed, ATP-dependent motor protein^{1,2} that moves along microtubules in discrete steps³ of 8 nm. *In vitro*, single molecules produce processive movement^{4,5}; motors typically take ~100 steps before releasing from a microtubule^{5–7}. A central question relates to mechanochemical coupling in this enzyme: how many molecules of ATP are consumed per step? For the actomyosin system, experimental approaches to this issue have generated considerable controversy^{8,9}. Here we take advantage of the processivity of kinesin to determine the coupling ratio without recourse to direct measurements of ATPase activity, which are subject to large experimental uncertainties^{8,10–12}. Beads carrying single molecules of kinesin moving on microtubules were tracked with high spatial and temporal resolution by interferometry^{3,13}. Statistical analysis of the intervals between steps at limiting ATP, and studies of fluctuations in motor speed as a function of ATP concentration^{14,15}, allow the coupling ratio to be determined. At near-zero load, kinesin molecules hydrolyse a single ATP molecule per 8-nm advance. This finding excludes various one-to-many and many-to-one coupling schemes, analogous to those advanced for myosin, and places severe constraints on models for movement.

Silica beads (0.5 µm diameter) were prepared with fewer than one molecule of kinesin, on average, bound nonspecifically to their surfaces. Beads were suspended in buffers containing variable amounts of ATP and introduced into a microscope flow cell. Individual diffusing beads were captured by an optical trap, deposited onto immobilized microtubules bound to the coverglass, and subsequent movements recorded with subnanometre resolution by optical-trapping interferometry^{3,13}. To ensure that beads were propelled by single molecules, they were prepared using extremely low concentrations of kinesin protein, sufficiently diluted from stock to make it unlikely that more than one molecule was present on a bead^{5,13}. In this regime, the fraction of beads moving as a function of kinesin concentration obeys Poisson statistics, confirming earlier findings^{4,5} (Fig. 1). To work at the lowest possible mechanical loads, the optical trap stiffness was set to just 7 fN nm⁻¹, producing a mean applied force of <0.9 pN in the optical trap (<15% of kinesin stall force¹³).

Average rates of kinesin movement were measured over three decades of ATP concentration. At saturating ATP levels, speeds were statistically identical to those measured by video tracking¹³ under no load, confirming the near-zero load condition. Speed data were well fit by Michaelis–Menten kinetics, with an apparent K_m for movement of $62 \pm 5 \mu\text{M}$ and k_{cat} of $680 \pm 31 \text{ nm s}^{-1}$, comparable to values found in microtubule gliding assays⁴ as well as solution studies of kinesin ATPase activity^{16–18}. This implies a Hill coefficient of one, and excludes models for movement that depart significantly from Michaelis–Menten kinetics¹⁹. Because hydrolysis activity in solution^{16–18} and speed *in vitro* display the same functional dependence on ATP, the coupling ratio—defined as the number of ATP molecules hydrolysed per mechanical advance—must be independent of ATP concentration. At limiting concentrations of ATP, k_{cat}/K_m gives a velocity of $11 \pm 1 \text{ nm s}^{-1} \mu\text{M}^{-1}$. In this regime, kinesin molecules advanced in clear increments of 8 nm (Fig. 2a), implying a stepping rate of $1.4 \pm 0.1 \mu\text{M}^{-1} \text{ s}^{-1}$ (Fig. 1). Statistical analysis of records, based on power spectra derived from pairwise distance differences³, revealed no signals corresponding to steps of

another size, although a small minority of steps of other sizes cannot be excluded (Fig. 2b,c). However, we did not confirm data from one study reporting roughly comparable numbers of 3-, 5- and 8-nm steps (ref. 20).

How long are the intervals between steps? Knowledge of this distribution at rate-limiting ATP may be used to determine the coupling. If kinesin molecules require just one ATP molecule per step, the distribution will be exponential, reflecting a solitary, rate-limiting biochemical reaction (binding of ATP). Conversely, a requirement for several (n) ATP molecules would lead to a distribution that is the convolution of n exponentials¹⁵. Pooling data from runs at 0.75, 1 and 2 μM ATP, during which steps could be identified in the records (Fig. 2), produced a distribution that was well fit by a single exponential (reduced $\chi^2 = 0.6$) with a rate of $1.1 \pm 0.1 \mu\text{M}^{-1} \text{s}^{-1}$ (Fig. 3a). This rate is in satisfactory agreement with the value derived independently from speed measurements (Fig. 1). In contrast, a convolution of two exponentials did not fit the data as well ($\chi^2 = 1.4$, $P(F = \chi^2_1/\chi^2_2) < 0.22$), and yielded a stepping rate ($0.8 \pm 0.06 \mu\text{M}^{-1} \text{s}^{-1}$) less compatible with k_{cat}/K_m . The observed step distribution supports the view that kinesin proteins bind a single ATP molecule before advancing by 8 nm.

To extend and confirm this finding, but without the need to identify individual stepwise transitions, we performed a fluctuation analysis of kinesin movement. For single processive motors, fluctuations about the average speed reflect the underlying enzyme stochasticity^{14,15}. Fluctuation analysis has been applied successfully to studies of the bacterial rotary motor²¹. Earlier work has shown that fluctuations can reveal the number of rate-limiting biochemical transitions per mechanical step^{14,15}; for kinesin molecules moving at saturating levels of ATP, there are approximately two such rate-limiting events¹⁴.

Fluctuation analysis of processive motor proteins relies on the determination of the randomness parameter, r , a dimensionless measure of the temporal irregularity between steps^{14,15}. A motor with $r = 0$ is perfectly clock-like, and takes an invariant time between steps. Conversely, a 'Poisson motor' with a single rate-limiting transition has $r = 1$. Additional rate-limiting transitions generally lead to $0 < r < 1$. For motors that step by a distance, d , whose positions are functions of time, $x(t)$, the randomness

parameter is defined as

$$r = \lim_{t \rightarrow \infty} \frac{\langle x^2(t) \rangle - \langle x(t) \rangle^2}{d \langle x(t) \rangle}, \quad (1)$$

where angle brackets denote the ensemble average. Both numerator and denominator in eqn (1) increase linearly with time (Fig. 2d), and their ratio approaches a constant whose reciprocal, r^{-1} , supplies a continuous measure of the number of rate-limiting transitions per step. Computations of the randomness parameter are robust to sources of thermal and instrumentation noise^{14,15}, thereby avoiding known difficulties associated with identifying individual steps against noisy backgrounds²². Nevertheless, the randomness parameter can be affected by factors that increase variance at long time scales, such as occasional backward steps, inactivated states, or steps of larger than normal size, which can all raise r beyond unity.

To ensure accurate determination of r , controls used both real and simulated data. We simulated runs at both saturating and limiting ATP conditions by generating stochastic 'staircase' records, consisting of 8-nm steps at either high or low average speeds ($\sim 630 \text{ nm s}^{-1}$, $\sim 10 \text{ nm s}^{-1}$) corrupted by gaussian white noise with total power comparable to that of real thermal noise. The times between steps were described by either an exponential distribution or a convolution of two exponentials. Fluctuation analysis of such records demonstrated

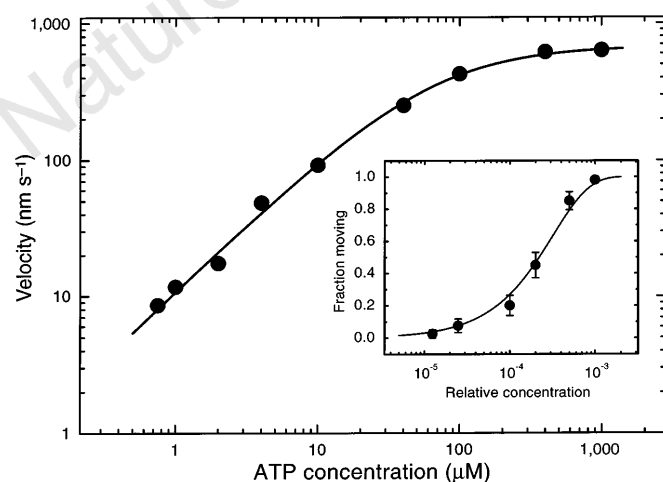


Figure 1 Average bead velocity, v , versus ATP concentration (double logarithmic plot). Error bars are smaller than data points (solid dots); each represents an average of 24–97 runs. Data fit to Michaelis-Menten kinetics (solid line), $v = k_{\text{cat}}[\text{ATP}]/(K_m + [\text{ATP}])$; $k_{\text{cat}} = 680 \pm 31 \text{ nm s}^{-1}$; apparent $K_m = 62 \pm 5 \mu\text{M}$. Inset, fraction of beads that moved, f , versus concentration of kinesin protein relative to stock, C (semilogarithmic plot). Values are expressed as mean $\pm$ $[f(1-f)/N]$. The one-parameter (λ) fit to Poisson statistics, $f = 1 - \exp(-\lambda C)$, (solid line, reduced $\chi^2 = 0.7$) confirms that single molecules suffice to move beads.

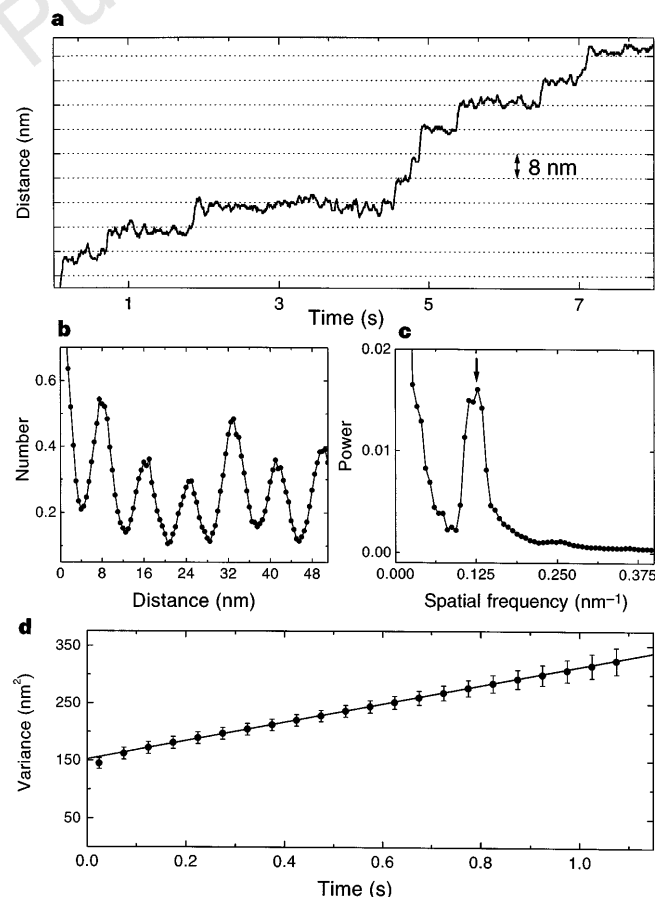


Figure 2 a, Sample record of movement at 2 μM ATP, showing the elementary steps (solid line). Horizontal grid lines (dotted lines) are spaced 8 nm apart. Data were median-filtered with a window width of 60 ms. b, Normalized histogram of pairwise distances between all pairs of data points in this record, showing a clear 8-nm periodicity. c, Normalized power spectrum of the data in b, displaying a single prominent peak at the reciprocal of 8 nm (arrow). d, Variance in position, averaged over 28 runs at 2 μM ATP (dots), and line fit over the interval 3.5 ms to 1.1 s. The y-intercept of the fit is determined by equipartition, $\langle x^2 \rangle = kT/\alpha$, where α is the combined stiffness of the optical trap and bead-microtubule linkage. The rapid rise in variance at short times reflects the brownian correlation time for bead position.

that r could be determined to better than 15%, and that the appropriate stepping statistics could thereby be distinguished (Table 1). To mimic data at low ATP, we fixed kinesin-coated beads on microtubules using 2 mM AMP-PNP, a non-hydrolysable ATP analogue. The microscope stage was then moved in nanometre-sized increments to generate simulated records with average speeds comparable to those in 1 μ M ATP (~ 10 nm s⁻¹) (Fig. 1). Once again, the number of rate-limiting transitions per step could readily be distinguished from the value of r . Individual steps in the latter records could also be detected by low-pass filtering the data. As a final control, we reconstructed interval distributions from these records; these were consistently fit by the appropriate model with either one or two rate-limiting steps (data not shown).

From records of kinesin-driven movement, we computed r as a function of ATP over a thousandfold range of concentration (Fig. 3). The qualitative shape of this relation had been predicted from the catalytic pathway^{14,15}. At high ATP levels, the curve is asymptotic to $r \approx \frac{1}{2}$. At intermediate ATP levels, the curve dips slightly, then rises at limiting ATP to a value near unity. The experimental value at saturating ATP confirms an earlier measurement¹⁴ and implies that each step involves a minimum of two rate-limiting transitions. Biochemical pathways derived from transient and steady-state kinetic studies also predict $r \approx \frac{1}{2}$ at high ATP levels, with the additional assumption that each mechanical step corresponds to one hydrolysis^{15,17,18}. For ATP levels near K_m , r drops as the rate of binding becomes comparably slow to other reactions. At limiting ATP, r rises through 1, reflecting a single rate-limiting transition occurring once per advance (ATP binding). The slight increase of r beyond unity reflects additional aspects of motor behaviour that tend to increase the variance.

Fluctuation analysis is robust to heterogeneity in the ATP binding rate (yielding heterogeneous mean speeds at limiting ATP), and controls allowed us to eliminate heterogeneity in bead size or in the stiffness of the bead–kinesin linkage as possible sources of additional variance (Table 1; Methods). Futile hydrolysis events that do not lead to a step can increase r , but not above unity¹⁴. Occasional pauses in movement, perhaps resulting from momentary sticking of beads to the coverglass or from motors transiently entering non-motile biochemical states, also increase variance. Sticking can be eliminated as an explanation because it would produce a concomitant reduction in brownian noise, and this was not observed. Transient inactive states causing $r \geq 1$ at limiting ATP result in even further increases beyond this value of randomness at moderate to saturating levels of ATP, and are therefore inconsistent with the

data. Two more possibilities merit examination: that ATP hydrolysis might occasionally produce backwards movement, a behaviour noted previously at higher external loads¹³ and also in this study; and that a single ATP hydrolysis might occasionally produce a double step of 16 nm, a transition that would appear indistinguishable from two 8-nm steps in rapid succession.

How frequent must such events be to give $r = 1.2$ –1.5 at limiting ATP? We addressed this question by using analytical expressions for r subject to each of the candidate mechanisms (Methods). The observed randomness could be generated if either $\sim 16\%$ of steps were 16 nm or $\sim 7\%$ of steps were backwards. By visual examination of records, we scored 6–9% of all steps at low ATP as being backwards. However, using the identical approach, 4–5% of steps were spuriously identified as backwards in controls in which beads were attached in rigor and the stage stepped stochastically in one direction only. Neither mechanism (nor admixtures of the two) can therefore be eliminated, and either produces an acceptable, one-parameter fit to the data (Fig. 3b). However, the additional variance derives from a minority of events: 84–93% of 8-nm steps correspond to a single ATP hydrolysis. In contrast, if 16-nm steps per hydrolysis were the norm, this behaviour would be unmistakable in the records, and the value of r would be 2 (or more) at limiting ATP. Our data are equally incompatible with 8-nm steps arising from the hydrolysis of two ATP molecules, because neither mechanism is adequate to raise the limiting value of r at low ATP from 0.5 to the observed value: to do so would require that we fail to detect that in excess of $\sim 20\%$ of steps went backwards, and no fraction whatsoever of 16-nm advances alone could raise r from 0.5 to beyond 1.

These findings place important constraints on theories of kinesin movement. The Michaelis–Menten dependence of kinesin velocity⁴ (Fig. 1) is consistent with a solitary ATP binding event per step, and is incompatible with a simultaneous requirement for two ATP molecules. However, Michaelis–Menten kinetics are equally consistent with kinetic schemes in which two ATP molecules bind sequentially to each of two heads. This can happen, in particular, whenever the two binding events are separated by a kinetically irreversible transition. Specific examples include ATP binding to each of two cooperative sites in an ordered sequence to produce a step, or ATP binding irreversibly in random order to each of two non-interacting sites. Such possibilities are excluded by our data. Furthermore, we can exclude models in which kinesin molecules hydrolyse multiple ATPs per 8-nm step²³, and all models in which ATP hydrolysis results in an average advance in excess of 8–9 nm, (the so-called ‘many-to-one’ coupling scenarios^{9,11}). Models that remain consistent with our study include those in which the centroid of the molecule advances in increments of 8 nm (refs 22, 24, 25), for example through alternating 16-nm steps by each of the two heads^{23,26,27}, or those in which each ATP molecule produces a composite of two shorter ‘substeps’²⁸. However, to be consistent with our findings, substeps must obey certain constraints. First, one of the substeps must be ATP-dependent, whereas the other must be ATP-independent. Second, the ATP-independent substep must be at least as fast as k_{cat} (or faster), in view of the behaviour at saturating ATP levels. Putting in numbers, this implies the maximum duration between substeps must be under ~ 15 ms at low loads, and so could not be resolved with instrumentation such as ours. If substeps were somehow to be detected under other conditions (for example, if 8-nm steps were actually a composite of 3- and 5-nm substeps that could be visualized at moderate load, separated by ~ 50 ms and more²⁰) this would imply that the ATP-independent substep was strongly load dependent, and that its rate slowed dramatically with increased load. However, one model that incorporates substeps predicts that the second substep should proceed rapidly, not slowly, under such loads²³. Addressing this issue experimentally will require additional work at higher forces.

The size of the kinesin motor domain²⁹ is quite small, $4.5 \times 4.5 \times 7.0$ nm. A sequence of ~ 40 amino acids beyond the

Table 1 Controls in velocity and randomness determination

Number of runs	Step size (nm)	Transitions per step	Nominal velocity (nm s ⁻¹)	Velocity (nm s ⁻¹)	Nominal r	r
Computer						
80	8	1	630	642 ± 16	1	1.12 ± 0.03
80	8	2	630	638 ± 4	0.5	0.54 ± 0.01
30	8	1	8	8.6 ± 0.5	1	1.02 ± 0.16
30	8	2	8	8.4 ± 0.2	0.5	0.49 ± 0.05
30	4	1	8	7.7 ± 0.3	1	1.01 ± 0.09
Stage						
25	8.6 ± 0.1	1	8.6 ± 0.1	9.0 ± 0.6	1	0.88 ± 0.09
23	5.1 ± 0.1	1	10.2 ± 0.2	9.8 ± 0.3	1	0.88 ± 0.09
27	9.4 ± 0.2	2	9.4 ± 0.2	9.5 ± 0.3	0.5	0.54 ± 0.07
30	4.8 ± 0.2	2	9.6 ± 0.4	9.8 ± 0.3	0.5	0.58 ± 0.06

Controls for determination of velocity and randomness parameter, expressed as mean $\pm$ s.e.m. Computer-generated simulations of displacement records were used to test analysis procedures at both fast and slow speeds, corresponding to saturating and limiting ATP concentrations, respectively. At slow speeds, physical controls were performed using kinesin-coated beads fixed to microtubules by AMP-PNP. To simulate movement, the position of the piezoelectric microscope stage was displaced stochastically in discrete increments, separated by time intervals picked at random from either an exponential distribution or a convolution of two exponentials, under computer control. The size of steps was determined separately, using beads stuck directly to the coverglass, thereby removing the dominant source of brownian noise in position measurement. In every case, experimental and nominal values were in satisfactory agreement, and the value of r returned by this analysis distinguished unambiguously between one or two transitions per step.

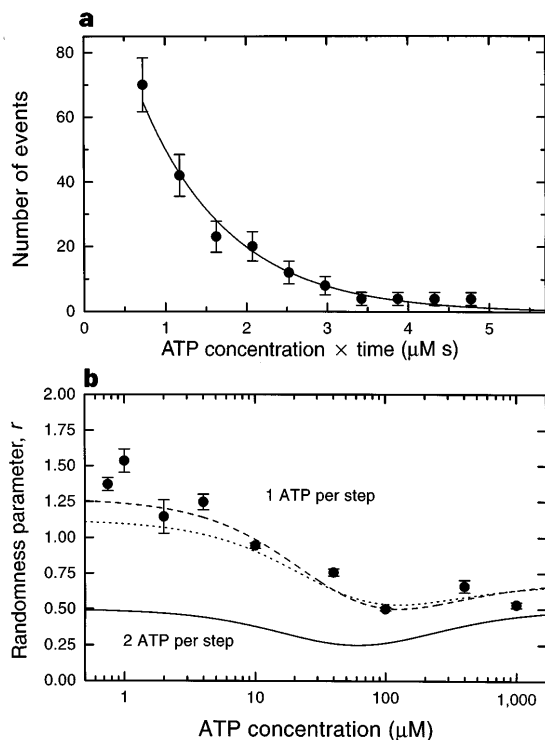


Figure 3 **a**, Histogram of times between steps for records at limiting ATP concentration (0.75, 1 and 2 μM). Data were pooled by weighting times by the ATP concentration. Values are expressed as $N \pm \sqrt{N}$. They fit to an exponential distribution (line) with a stepping rate of $1.1 \pm 0.1 \mu\text{M}^{-1} \text{s}^{-1}$ (reduced $\chi^2 = 0.6$). **b**, Semilogarithmic plot of the randomness parameter, r , versus ATP concentration. Experimental data (dots) and associated errors (s.e.m., based on 24–97 runs per determination) are compared to various models. Computed value of r , based on the k_{cat} and K_m values from Fig. 1, assuming two rate-limiting biochemical transitions at saturating ATP concentrations and two hydrolyses per step (solid line). There are no free parameters. Computed one-parameter (P_-) fit for r , based on the assumption of two rate-limiting biochemical transitions at saturating ATP concentrations and a single ATP hydrolysis per step, plus $P_- = 7\%$ backwards steps (dotted line). Computed one-parameter fit for r , based on the same assumptions, plus $P_{16} = 16\%$ of 16-nm steps (dashed line).

carboxy end of the crystal structure is thought to form an α -helix, and may supply an additional $\sim 3 \text{ nm}$ of leeway before the two kinesin heavy chains joint to form a coiled-coil tail. However, this segment is probably too short to constitute a 'lever arm' analogous to the structure proposed to be the site of displacement and force generation in myosin³⁰. An 8-nm step therefore poses challenges for understanding the molecular basis of kinesin movement. We anticipate that further applications of fluctuation analysis will supply useful clues to unravelling the mechanochemistry. Moreover, the power of this approach to provide insight into single-molecule kinetics makes it applicable to other processive systems, such as nucleic acid enzymes, including polymerases, helicases, topoisomerases and ribosomes. □

Methods

Assays. Motility assays were performed as previously described^{3,5,13}, except that microtubules were made from bovine brain tubulin (Cytoskeleton), coverslips were coated with poly-L-lysine rather than silanized, KCl was replaced by K-acetate, and an oxygen scavenging system (250 $\mu\text{g ml}^{-1}$ glucose oxidase, 30 $\mu\text{g ml}^{-1}$ catalase, 4.5 mg ml^{-1} glucose; Sigma) was used. An ATP regenerating system stabilized low ATP levels^{5,13}. Kinesin was purified from squid optic lobe as described¹. As a precaution against variability and to ensure work in the single molecule regime, only data from assays in which fewer than half of all beads moved were analysed.

Calibration and analysis. Optical-trapping interferometry and calibrations were performed as described^{3,13}. A trap stiffness of 7 fN nm^{-1} (average load $< 0.9 \text{ pN}$ over 50–200 nm from the trap centre) was obtained using 12 mW illumination at the specimen plane from a Nd:YLF laser (1047 nm, Spectra Physics). Interferometer output signals were sampled at 2 kHz, anti-alias filtered at 1 kHz, digitized at 12 bits, and stored by computer. Analysis programs were written in LabView 4.01 (National Instruments). Velocities, fluctuations and pairwise distances were analysed as described previously^{3,14}. Multiple runs over 50–200 nm from the trap centre were corrected to adjust for the kinesin-to-bead linkage compliance^{3,13}, except for the example in Fig. 2a, which was selected for low noise³. Mean velocities at each ATP concentration, $\langle v \rangle = \langle x(t) \rangle / t$, were determined by line fits to the average pairwise separation versus time for each run; slopes from such fits were averaged over all runs. The variance, $\langle x^2(t) \rangle - \langle x(t) \rangle^2$, was computed as the mean-squared deviation from the trajectory $\langle x(t) \rangle = \langle v \rangle t$, time-averaged for each run, then averaged over all runs. A line fit was made to this variance over the first 20 nm/ $\langle v \rangle$ time interval, after excluding the first 3.5 ms, because of the ~ 1 –3 ms correlation time for bead position^{3,14}. Theory and simulations show that a linear increase of variance with time (Fig. 2d) reflects a lack of any significant heterogeneity in mean velocity. The latter would produce a variance that grew quadratically with time. The standard error in variance also grew linearly with time, thereby supplying an error estimate for the slope. Randomness, r , was computed from the slope of the variance line fit divided by $d\langle v \rangle$. Experimental and control runs at limiting ATP levels were smoothed over a 5-ms time window with a linear filter, then decimated at 5-ms intervals, a procedure that reduces computation time but does not affect r . Subject to the minimal assumption of exactly two rate-limiting transitions per step at saturating ATP, r as a function of ATP concentration is determined by two parameters: the apparent Michaelis–Menten constant, and the ATP binding rate, determined in Fig. 1. To ascertain the effect of backward movements or 16-nm steps on r , we used the following expressions: $r = 1/(1 - 2P_-) + (\bar{r} - 1)/(1 - 2P_-)$, and $r = (1 + 3P_{16})/(1 + P_{16}) + (\bar{r} - 1)/(1 + P_{16})$, where P_- and P_{16} are the probabilities of backwards and 16-nm events, respectively, and $\bar{r}$ is the randomness in their absence (M.J.S. and S.M.B., unpublished). By using experimental values for K_m and binding rate, the randomness data of Fig. 3 were fit to each of these expressions using a single free parameter. For analysis of pairwise distances and times between advances, runs at 0.75, 1 and 2 μM ATP were median-filtered with a window of 20–75 ms. Step times were identified by eye using custom software that returned the interval selected by cursors superposed on a graphical display of displacement versus time. Events shorter than 0.5 $\mu\text{M s}$ were excluded from analysis to avoid artefacts arising from the missing-event problem¹⁴.

Received 13 February; accepted 12 May 1997.

- Vale, R. D., Reese, T. S. & Sheetz, M. P. Identification of a novel force-generating enzyme, kinesin, involved in microtubule-based motility. *Cell* **42**, 39–50 (1985).
- Kreis, T. & Vale, R. D. *Guidebook to the Cytoskeletal and Motor Proteins* (Oxford Univ. Press, 1993).
- Svoboda, K., Schmidt, C. F., Schnapp, B. J. & Block, S. M. Direct observation of kinesin stepping by optical trapping interferometry. *Nature* **365**, 721–727 (1993).
- Howard, J., Hudspeth, A. J. & Vale, R. D. Movement of microtubules by single kinesin molecules. *Nature* **342**, 154–158 (1989).
- Block, S. M., Goldstein, L. S. B. & Schnapp, B. J. Bead movement by single kinesin molecules studied with optical tweezers. *Nature* **348**, 348–352 (1990).
- Vale, R. D. *et al.* Direct observation of single kinesin molecules moving along microtubules. *Nature* **380**, 451–453 (1996).
- Hackney, D. D. Highly processive microtubule-stimulated ATP hydrolysis by dimeric kinesin head domains. *Nature* **377**, 448–450 (1995).
- Burton, K. Myosin step size: estimates from motility assays and shortening muscle. *J. Muscle Res. Cell Motil.* **13**, 590–607 (1992).
- Yanagida, T., Harada, Y. & Ishijima, A. Nano-manipulation of actomyosin molecular motors *in vitro*: a new working principle. *Trends Biochem. Sci.* **18**, 319–324 (1993).
- Sowerby, A. J., Sehara, C. K., Lee, M. & Bagshaw, C. R. Turnover of fluorescent nucleoside triphosphates by isolated immobilized myosin filaments: transient kinetics on the zeptomole scale. *J. Mol. Biol.* **234**, 114–123 (1993).
- Harada, Y., Sakurada, K., Aoki, T., Thomas, D. D. & Yanagida, T. Mechanochemical coupling in actomyosin energy transduction studied by *in vitro* movement assay. *J. Mol. Biol.* **216**, 49–68 (1990).
- Toyoshima, Y. Y., Kron, S. J. & Spudis, J. A. The myosin step size: measurement of the unit displacement per ATP hydrolyzed in an *in vitro* assay. *Proc. Natl Acad. Sci. USA* **87**, 7130–7134 (1990).
- Svoboda, K. & Block, S. M. Force and velocity measured for single kinesin molecules. *Cell* **77**, 773–784 (1994).
- Svoboda, K., Mitra, P. P. & Block, S. M. Fluctuation analysis of motor protein movement and single enzyme kinetics. *Proc. Natl Acad. Sci. USA* **91**, 11782–11786 (1994).
- Schnitzer, M. J. & Block, S. M. Statistical kinetics of processive enzymes. *Cold Spring Harb. Symp. Quant. Biol.* **60**, 793–802 (1995).
- Hackney, D. D. The rate-limiting step in microtubule-stimulated ATP hydrolysis by dimeric kinesin head domains occurs while bound to the microtubule. *J. Biol. Chem.* **269**, 16508–16511 (1994).

17. Gilbert, S. P., Webb, M. R., Brune, M. & Johnson, K. A. Pathway of processive ATP hydrolysis by kinesin. *Nature* **373**, 671–676 (1995).
18. Ma, Y. & Taylor, E. W. Kinetic mechanism of the kinesin motor domain. *Biochemistry* **34**, 13242–13251 (1995).
19. Leibler, S. & Huse, D. A. Porters versus rowers: a unified stochastic model of motor proteins. *J. Cell Biol.* **121**, 1357–1368 (1993).
20. Coppin, C. M., Finer, J. T., Spudich, J. A. & Vale, R. D. Detection of sub-8-nm movements of kinesin by high-resolution optical-trap microscopy. *Proc. Natl Acad. Sci. USA* **93**, 1913–1917 (1996).
21. Samuel, A. D. T. & Berg, H. C. Torque-generating units of bacterial flagellar motor step independently. *Biophys. J.* **71**, 918–923 (1996).
22. Block, S. M. & Svoboda, K. Analysis of high resolution recordings of motor movement. *Biophys. J.* **68**, 230s–241s (1995).
23. Duke, T. & Leibler, S. Motor protein mechanics: a stochastic model with minimal mechanochemical coupling. *Biophys. J.* **71**, 1235–1247 (1996).
24. Hackney, D. D. Evidence for alternating head catalysis by kinesin during microtubule-stimulated ATP hydrolysis. *Proc. Natl Acad. Sci. USA* **91**, 6865–6869 (1994).
25. Gelles, J. *et al.* Structural and functional features of one- and two-headed biotinylated kinesin derivatives. *Biophys. J.* **68**, 276s–282s (1995).
26. Howard, J. The movement of kinesin along microtubules. *Annu. Rev. Physiol.* **58**, 703–729 (1996).
27. Peskin, C. S. & Oster, G. Coordinated hydrolysis explains the mechanical behavior of kinesin. *Biophys. J.* **68**, 202s–211s (1995).
28. Derenyi, I. & Vicsek, T. The kinesin walk: a dynamic model with elastically coupled heads. *Proc. Natl Acad. Sci. USA* **93**, 6775–6779 (1996).
29. Kull, F. J., Sablin, E. P., Lau, R., Fletterick, R. J. & Vale, R. D. Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature* **380**, 550–555 (1996).
30. Block, S. M. Fifty ways to love your lever: myosin motors. *Cell* **87**, 151–157 (1996).

Acknowledgements. We thank S. Gross, W. Ryu, L. Satterwhite, M. Wang and especially K. Visscher for technical assistance and discussions; S. Gross for help in purifying kinesin; and P. Mitra and K. Svoboda for advice in the early stages of this project. This work was supported by a grant from NIGMS (S.M.B.) and predoctoral fellowships from NSF and American Heart Association (M.J.S.).

Correspondence should be addressed to M.J.S. (e-mail: schnitzr@princeton.edu).

Coupling of kinesin steps to ATP hydrolysis

Wei Hua*, Edgar C. Young, Margaret L. Fleming & Jeff Gelles

* Biophysics and Structural Biology Graduate Program, Department of Biochemistry, and The Center for Complex Systems, Brandeis University, Waltham, Massachusetts 02254, USA

A key goal in the study of the function of ATP-driven motor enzymes is to quantify the movement produced from consumption of one ATP molecule^{1–3}. Discrete displacements of the processive motor kinesin along a microtubule have been reported as 5 and/or 8 nm (refs 4, 5). However, analysis of nanometre-scale movements is hindered by superimposed brownian motion. Moreover, because kinesin is processive and turns over stochastically, some observed displacements must arise from summation of smaller movements that are too closely spaced in time to be resolved. To address both of these problems, we used light microscopy instrumentation⁶ with low positional drift (< 39 pm s⁻¹) to observe single molecules of a kinesin derivative moving slowly (~2.5 nm s⁻¹) at very low (150 nM) ATP concentration, so that ATP-induced displacements were widely spaced in time. This allowed increased time-averaging to suppress brownian noise (without application of external force^{4,5}), permitting objective measurement of the distribution of all observed displacement sizes. The distribution was analysed with a statistics-based method which explicitly takes into account the occurrence of unresolved movements, and determines both the underlying step size and the coupling of steps to ATP hydrolytic events. Our data support a fundamental enzymatic cycle for kinesin in which hydrolysis of a single ATP molecule is coupled to a step distance of the microtubule protofilament lattice spacing of 8.12 nm (ref. 7). Step distances other than 8 nm are excluded, as is the coupling of each step to two or more consecutive ATP hydrolysis reactions with similar rates, or the coupling of two 8-nm steps to a single hydrolysis. The measured ratio of ATP consumption rate to stepping rate is invariant over a wide range of ATP concentration, suggesting that the 1 ATP to 8 nm coupling inferred from behaviour at low ATP can be generalized to high ATP.

Box 1 Distribution of displacements observed with limited time resolution

Detection of processive motor steps can be modelled analytically. We assume that, in the absence of noise, every displacement detected has size nd_1 , where n is a positive integer and d_1 is the unitary step size. We model the effects of a step discrimination method in which the time interval t between neighbouring unitary steps is detected only if $t > t_r$ the time resolution, t_r . In this case, the probability that $n = 1$ is

$$P_1 = P(t > t_r),$$

where $P(t > t_r)$ is the probability that the interval $t > t_r$. The probability that $n = 2$ is the probability of not detecting the interval between one pair of unitary steps and detecting the interval between the next pair:

$$P_2 = [1 - P(t > t_r)]P(t > t_r) = (1 - P_1)P_1.$$

In general,

$$P_n = P_1(1 - P_1)^{n-1}.$$

The histogram of observed displacements will be:

$$H(x) = N' \begin{cases} P_n & x = nd_1 \text{ with } n = 1, 2, 3, \dots \\ 0 & \text{elsewhere,} \end{cases}$$

where N' is the total number of displacements detected. N' also corresponds to the maximum value in the cumulative histogram,

$$C(d) = \int_0^d H(y) dy.$$

$C(d)$ differs from the cumulative histograms shown in Fig. 3, in that the latter include effects of experimental noise (see Methods). The total number of unitary steps $\tilde{N} = D/d_1$, where D is the total distance moved, and

$$\tilde{N} = N' \sum_{i=1}^{\infty} iP_i = N'/P_1.$$

At limiting ATP concentrations, mechanisms with the same d_1 but different coupling of steps to ATP hydrolysis give different $P(t > t_r)$ and therefore different histograms. For example, for a mechanism in which each unitary step is coupled to the hydrolysis of one ATP molecule,

$$P(t > t_r) = e^{-t_r/\tau},$$

where $\tau = Td_1/D$ is the average time interval between steps, and T is the total time. In contrast, if each step is followed by two consecutive ATP hydrolysis events with rate constant ratio k ,

$$P(t > t_r) = \frac{e^{-[(k+1)t_r/\tau]} - k e^{-[k+1)t_r/\tau]}{1 - k}.$$

These results can be generalized to apply to mechanisms that include more than one unitary step size.

K448-BIO is a recombinant biotinylated protein containing the amino-terminal 448 amino acids of the *Drosophila* kinesin heavy chain (including the motor domain, which is sufficient for microtubule-based motility⁸). It forms stable dimers⁹ and so, like intact kinesin¹⁰, it contains two active sites. K448-BIO displays kinesin-like ATPase activity⁹, and single dimers of K448-BIO attached to streptavidin beads move processively^{11,12} (that is, for hundreds of nanometres without detaching from the microtubule), like intact kinesin^{13,14} and other dimeric kinesin derivatives¹⁵. The rate of forward translocation by single molecules displays hyperbolic dependence on ATP concentration (Fig. 1a). The fitted Michaelis–Menten parameters from velocities at ATP concentrations $\geq 5 \mu\text{M}$ are confirmed by the slope of the linear ATP dependence of velocities at sub-micromolar ATP levels (Fig. 1a, inset). This strict hyperbolic dependence is consistent with a reaction sequence in which the two active sites of the enzyme bind ATP in a sequential manner, provided that an irreversible step, such as phosphate release, occurs between the two binding reactions. (A mechanism without an intervening irreversible step has a non-hyperbolic rate law¹⁶.)

The ATP consumption associated with the motor's movement

can be quantified by comparing its single-molecule translocation velocity with its microtubule-activated ATPase rate measured in solution conditions at which all motor molecules are translocating¹³. We measured K448-BIO ATPase rates at maximally activating microtubule concentration. Over an ATP concentration range of 3–1,000 μM , the ATPase rate changes in parallel with the movement velocity (Fig. 1a), so that the ‘coupling ratio’ of ATP hydrolysis to distance moved is constant within experimental error (Fig. 1b). The value of the ratio corresponds approximately to one ATP hydrolysis for each 8.12 nm of forward translocation. However, the assumption that in the solution ATPase experiments all molecules are actively translocating may not be wholly accurate, as a normal active motor might spend a fraction of time in a ‘quiescent’ non-translocating state. The ratio observed probably excludes the consumption of high numbers of ATP molecules per 8 nm moved, or movement over many lattice spacings per ATP molecule, but may not be precise enough to distinguish, for example, whether the true coupling ratio is two ATP molecules per 8 nm, one ATP per 8 nm, or one ATP per 16 nm.

To characterize the coupling of kinesin movement to ATP hydrolysis by a complementary approach, we used video-enhanced differential interference contrast microscopy to follow with nanometre-scale precision⁶ the movement of bead-labelled K448-BIO molecules on microtubules. K448-BIO moves in a series of discrete

displacements rather like those observed in previous studies of intact kinesin^{4,5}. At very low (150 nM) ATP concentrations, displacement events are well separated by dwell times (Fig. 2) and could be discriminated reliably by computer. The absence of applied force makes it unnecessary to correct displacement measurements for mechanical compliance^{4,5}, and eliminates concerns that applied load may alter the coupling mechanism². Because the underlying unitary translocations (‘steps’) are stochastic, we expect that some groups of neighbouring steps may not be resolved and will give rise to displacements that are sums of the unitary step sizes. The number of unresolved steps (see Box 1) is significant when the time resolution of the step discrimination algorithm (~ 0.6 s) is not much less than the mean time between steps (2.5 s).

To take unresolved steps into account, we examined the size distribution of all detected displacements. This is valuable because this distribution will in general have a different shape for movement mechanisms with different step size, or with different coupling between ATP hydrolytic events and steps (Box 1). Another distinguishing factor is N , the total number of positive displacements observed in a given total time (T) and total distance (D) moved. A cumulative histogram of the displacements is a simple way of examining both of these features: the histogram shape shows distinctive features of the displacement distribution, and N is the histogram asymptote. We numerically simulated distance–time

Figure 1 a, Velocities parallel to microtubule axis of beads of 100-nm diameter conjugated to single molecules of kinesin derivative K448-BIO. Points show mean velocities ($\pm$ s.d.) from 6–40 observations at each concentration of ATP (with the exception of 0.55 μM ATP, where only one observation was made). Line shows least-squares fit to hyperbolic ATP dependence of all data collected at ATP concentration $\geq 5 \mu\text{M}$; the fitted values ($\pm$ s.e.) for parameters were $V_{\text{max}} = 819 \pm 21 \text{ nm s}^{-1}$, $K_m = 46.4 \pm 4.4 \mu\text{M}$. Inset, mean velocities at $< 5 \mu\text{M}$ ATP (linear concentration scale). **b**, ATP consumed by K448-BIO dimers per movement step assuming each step is the microtubule lattice spacing of 8.12 nm (ref. 7). For each ATP concentration, ATPase rate in solution at saturating microtubule concentration was expressed as turnovers per second per exchangeable nucleotide site and multiplied by 2 (because a translocating dimer must have two exchangeable sites). Single-molecule movement velocity at the same ATP concentration was calculated from the fit parameters from **a** and divided by 8.12 nm to express the rate as steps per second. Points show the ratio of hydrolysis rate to stepping rate: the horizontal line indicates the mean value of 0.77 ATP molecules per dimer per step (s.d. = 0.24, $n = 20$).

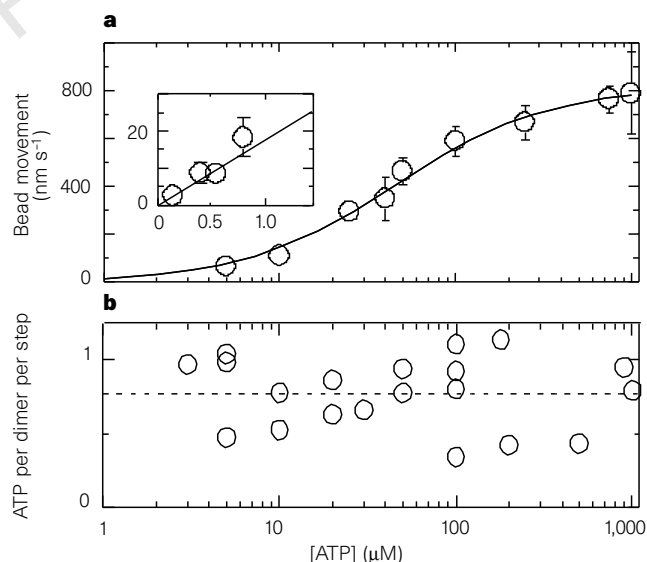
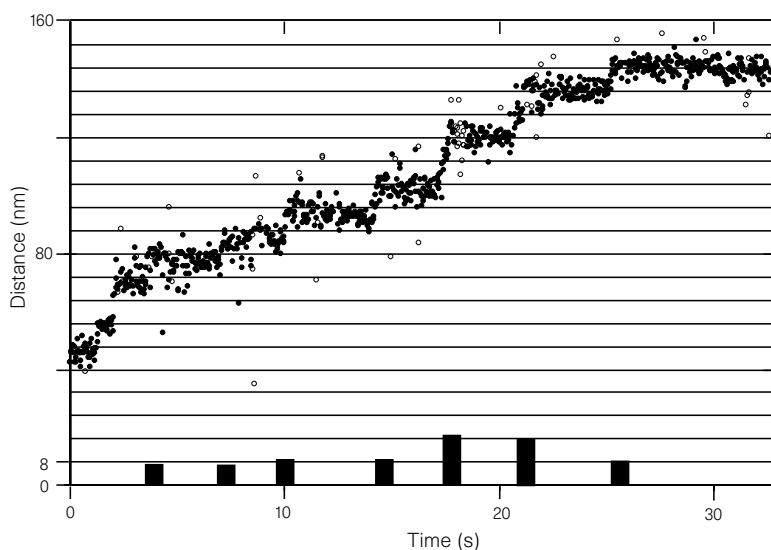


Figure 2 Movement of a bead-labelled single K448-BIO molecule along a microtubule in 150 nM ATP. Points mark the bead position on the microtubule axis; the time interval is 33 ms. Open circles represent measurements in which a free bead in solution diffused within one bead image diameter of the measured bead, possibly resulting in an inaccurate measurement. Bars indicate displacements detected in the same data record by an automated step-discrimination algorithm. The height of the bar represents the displacement magnitude, and its horizontal position represents the time of the event. The displacements near time 2 s were not detected by the algorithm because the first and last 1.98 s of the data record are excluded from the analysis (see Methods).



records for various coupling mechanisms and subjected them to the same step discrimination algorithm (that is, with the same time resolution for detection of displacements) as was used for experimental traces. The cumulative histogram of displacements from the simulation of each mechanism can be compared directly with the experimental histogram (Fig. 3). The experimental data are consistent with a simple mechanism for kinesin movement in which each ATP hydrolysis is coupled to a step from one tubulin dimer on the microtubule to the adjacent dimer along the same protofilament (the '8.12' mechanism), but not with several other proposed kinesin mechanisms that were consistent with our solution ATPase measurements.

Mechanisms with a single ATP hydrolytic event coupled to an average unitary step size of less than 8 nm will have a value of N (for a given D and T) higher than that in the '8.12' mechanism. This is illustrated in Fig. 3 by the '3-5-8' mechanism (random steps of 3, 5 and 8 nm with equal probabilities, as discussed in ref. 5, with each step coupled to a single hydrolytic event). The value of N is greater than that in the '8.12' mechanism because the average unitary step size (5.3 nm) is smaller. Furthermore, the shape of the displacement distribution is far from that of the experimental data, providing strong evidence against this mechanism. A similar mechanism with steps of 3 and 5 nm also fails to fit (Fig. 3).

The cumulative histograms also exclude mechanisms that postulate a constant step size of 8 nm but a coupling ratio different from that of the '8.12' mechanism. For the '1ATP/2step' mechanism, in which two steps of 8.12 nm are coupled to a single ATP hydrolytic event, the average time interval between the two steps (≤ 10 ms to produce the 100 s^{-1} mean stepping rate at saturating ATP) is much less than the time resolution. This yields a small N because the behaviour is essentially identical to stepping with unitary step size of 16.24 nm. The '2ATP/step' mechanism, with an 8.12 nm step coupled to a pair of ATP hydrolytic events with identical rates, has, compared with the '8.12' mechanism, a larger proportion of

time intervals between steps (Box 1) long enough to be resolved by the discrimination algorithm, and larger N . Mechanisms with one step coupled to more than two identical ATP hydrolysis reactions give a cumulative histogram (not shown) similar to that of the '2ATP/step' mechanism.

The data also exclude modified versions of the '2ATP/step' mechanism (not shown) in which one step is coupled to a pair of non-identical ATP hydrolysis reactions, unless the two productive ATP-binding events occur at rates that differ by a factor of >15 . This difference in rates can occur only if the enzyme is capable of two distinct modes of ATP hydrolysis. No evidence for this unusual possibility has been found in studies of kinesin kinetics^{17,18} so, although we do not discount it entirely, we consider it improbable.

The distribution of time intervals between observed displacements (not shown) is monoexponential, which is consistent with the '8.12' mechanism.

In summary, every displacement of 8 nm requires at least one ATP hydrolysis, and probably not more than one, so 8 nm is the distance associated with what can rightly be called a complete enzymatic cycle. In stable complexes of kinesin with a microtubule, kinesin interacts with only one tubulin dimer¹⁹⁻²². An 8-nm cycle can then be interpreted as reliable movement of kinesin from one tubulin dimer to the adjacent tubulin dimer of the same microtubule protofilament. This confirms the observation at high ATP concentration that kinesin molecules follow a single protofilament^{6,11,23}. We do not exclude the possibility that at some point in the cycle kinesin makes a step with an apparent displacement different from 8 nm (such as the 5-nm displacements previously reported⁵). Nevertheless, we predict that in our experiments such steps must be part of a series of displacements that sum to 8 nm with each iteration of the ~ 10 -ms cycle. Our data (temporal and spatial distribution of discrete movements, ATP dependence of movement velocity, and solution ATPase) are fully consistent with the tight coupling of ATP consumption to forward movement, and with the currently

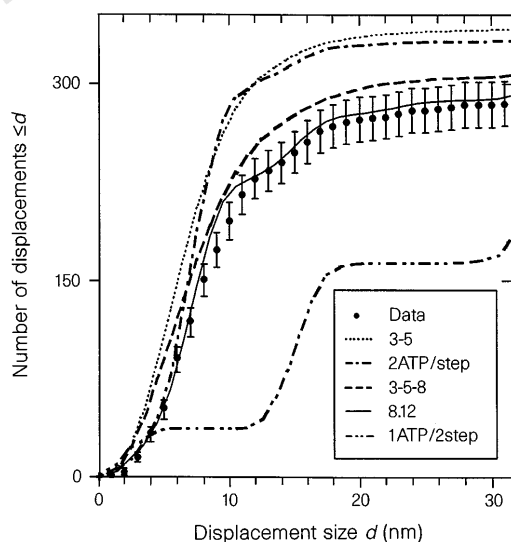


Figure 3 Cumulative histograms of displacements of bead-labelled enzyme compared to histograms calculated for various enzyme mechanisms. The number of positive displacements less than or equal to a displacement size d is plotted on the vertical axis. Points represent a histogram ($\pm$ s.e.) of $N = 294$ observed positive displacements (negative displacements^{4,5}, 26 of which were >5.1 nm, were not included) from experiments at 150 nM ATP, with total bead movement distance $D = 2,678$ nm and total bead movement duration $T = 1,082$ s. Lines indicate simulated behaviour of various enzyme mechanisms. Note that the simulations have no adjustable parameters and are not fit to the experimental data. Each line is the average of data from 20 simulations, each with D and T identical to the experimental data record. The mean time between ATP hydrolytic

events (t_{ATP}) was calculated from D and T and the properties of each mechanism. For the '8.12' mechanism, each hydrolytic event is coupled to a single step of 8.12 nm ($t_{\text{ATP}} = 3.28$ s, $N = 291$); '2ATP/step' mechanism, each 8.12 nm step is coupled to a pair of hydrolytic events ($t_{\text{ATP}} = 1.64$ s, $N = 332$); '1ATP/2step' mechanism, each hydrolytic event is coupled to a pair of 8.12-nm steps separated by 10 ms (8.12 nm divided by V_{max} ; $t_{\text{ATP}} = 6.56$ s, $N = 193$); '3-5' mechanism, random 3- and 5-nm steps, with each step coupled to a single ATP hydrolytic event ($t_{\text{ATP}} = 1.64$ s, $N = 341$); '3-5-8' mechanism, 3-, 5- and 8-nm steps randomly mixed with equal probabilities and with each step coupled to a single hydrolytic event ($t_{\text{ATP}} = 2.15$ s, $N = 307$). The '8.12' mechanism is clearly the closest match to the data.

favoured 'alternating sites' mechanisms for kinesin movement^{3,24}. The coupling ratio measured in the macroscopic steady-state experiments reflects its microscopic foundation, namely a regular 8-nm mechanochemical cycle that can be identified by direct observation of stochastic movements. □

Methods

Enzyme preparation and ATPase assay. K448-BIO was purified⁹, supplemented with 1 M sucrose, frozen in liquid nitrogen, and stored at -80 °C. Taxol-stabilized microtubules were prepared as described²⁵. ATPase assays were performed at 22–25 °C using radiolabelled nucleotide methods adapted from ref. 26: K448-BIO (0.2–0.7 µg ml⁻¹) was mixed with [α -³²P]ATP in reaction buffer (ABT buffer¹¹ supplemented with 2 mM dithiothreitol (DTT) and 0.1 mg ml⁻¹ bovine serum albumin); reactions were started by adding 1.7–2.0 mg ml⁻¹ microtubules. At each time point, a portion was quenched by a 5–6-fold dilution into ice-cold 1 M HCl spiked with unlabelled ATP and ADP standards; the mixture was neutralized with an equal volume of 1 M Tris base, filtered (Millipore Ultrafree-MC polysulphone, 30,000 NMWL), and chromatographed on TSK-GEL DEAE-2SW HPLC for quantification of labelled ATP and ADP. Corrections were made for impurities in the radiolabelled nucleotide stock²⁶. ATPase rates were constant up to 15% completion of reaction; the rate from each enzyme–microtubule mixture was corrected by subtraction of the rate for microtubules alone and expressed as enzyme turnovers per second per exchangeable nucleotide binding site. Concentration of exchangeable nucleotide sites in K448-BIO preparations was measured²⁷ by incubating 3–9 µg ml⁻¹ K448-BIO with 0.5–2 µM [α -³²P]ATP in reaction buffer supplemented with 0.1 mg ml⁻¹ creatine kinase and 2 mM phosphocreatine. Creatine-kinase-inaccessible [³²P]ADP (that is, K448-BIO exchangeable nucleotide sites) in quenched portions of this mixture was determined as above. Accumulation of [³²P]ADP displayed kinetics similar to those of [³²P]ADP depletion after addition of 170 µM unlabelled ATP.

Motility assays. Streptavidin beads were mixed with K448-BIO at a ratio of 0.2–1.0 enzyme dimers per bead¹¹, and motility was measured (at 22–25 °C) as described¹¹, except that samples were supplemented with 2 mM DTT and ATP concentration was varied. Samples with <5 µM ATP were supplemented with 7 µg ml⁻¹ creatine kinase and 2 mM phosphocreatine to prevent ATP depletion. As expected for beads driven by single kinesin molecules¹⁴, beads typically detached from the microtubule before reaching the microtubule end (mean travel distances, 300–1,000 nm at 5–1,000 µM ATP). In most experiments, bead velocity was calculated from positions of video cursors manually superimposed on the bead image in video frames at the start and end of the movement episode. Data for mean velocity v at [ATP] $\geq$ 5 µM were fitted to the Michaelis–Menten equation $v = V_{\max}[\text{ATP}] / ([\text{ATP}] + K_m)$ by least-squares regression (weighted by reciprocal of standard deviation). For nanometre-precision measurements of movement¹¹ at 0.15 or 0.8 µM ATP, carboxyl-polystyrene beads 100 nm in diameter were adsorbed to the coverslip surface (after microtubules) to serve as stationary reference points.

Nanometre-precision position measurements. We made simultaneous position measurements on multiple beads in the microscope field. Positions of moving beads were measured relative to the position of the coverslip determined from the averaged positions of $\geq$ 10 stationary beads, to compensate for mechanical drift in the microscope⁶. The residual drift rate was <39 pm s⁻¹ in both x and y in 8 independent measurements; precision⁶ was 2.0 nm in x and 0.8 nm in y . Position on the microtubule axis is calculated by projecting the y position to the measured¹¹ microtubule orientation, which was always <2° from the y direction. Beads that moved <15 nm were excluded from further analysis. A free bead in solution diffusing within one image diameter of the measured bead interfered with ~6% of the position measurements (Fig. 2).

Step discrimination algorithm. Axial position records were smoothed with a recursive median filter^{4,28} with rank 15 (495-ms window), the first derivative was calculated with a 14-point (462 ms) quadratic convolute²⁹, and the first and last 1.98 s of data were discarded. The beginning and end of bead dwell positions were taken as the crossings of a velocity threshold value of 6.1 nm s⁻¹. The magnitude of the displacement between adjacent dwells was calculated as the difference in mean dwell positions, with positive direction defined as the direction of the net movement in the data record. The algorithm was validated by demonstrating that, for all enzyme mechanisms discussed, its

effect could be duplicated by a simple constant-parameter linear system (W.H. and J.G., unpublished results) in which those unitary steps separated by a time interval longer than a time resolution t_r are resolved, and those separated by a shorter interval are summed. This system shows the desirable properties that N is nearly independent of small changes in noise amplitude, but is approximately inversely proportional to unitary step size and stepping rate. For all mechanisms examined, N' (Box 1) changed in parallel with N .

Monte-Carlo simulations with added noise. To simulate the behaviour of different coupling mechanisms (Fig. 3), we generated traces each with displacements of the specified sizes separated by random time intervals chosen from an exponential or biexponential distribution. Data records (with standard deviations of 2.5, 3.3 and 3.6 nm; by comparison, the precision of the moving bead traces was ~3.8 nm) from control experiments in which 1 mM AMP-PNP was substituted for ATP were added to the generated traces to simulate experimental noise. Curves in Fig. 3 were produced by applying the step discrimination algorithm to traces. The variance of the cumulative histogram outputs from 20 independent simulations of the '8.12' mechanism is that expected from statistical variation alone, indicating that the discrimination algorithm does not introduce significant additional variance in the histograms.

Received 13 February; accepted 12 May 1997.

1. Taylor, E. W. Variations on the theme of movement. *Nature* **361**, 115–116 (1993).
2. Ishijima, A. et al. Single-molecule analysis of the actomyosin motor using nano-manipulation. *Biochem. Biophys. Res. Commun.* **199**, 1057–1063 (1994).
3. Hackney, D. D. Evidence for alternating head catalysis by kinesin during microtubule-stimulated ATP hydrolysis. *Proc. Natl Acad. Sci. USA* **91**, 6865–6869 (1994).
4. Svoboda, K., Schmidt, C. F., Schnapp, B. J. & Block, S. M. Direct observation of kinesin stepping by optical trapping interferometry. *Nature* **365**, 721–727 (1993).
5. Coppin, C. M., Finer, J. T., Spudis, J. A. & Vale, R. D. Detection of sub-8-nm movements of kinesin by high-resolution optical-trap microscopy. *Proc. Natl Acad. Sci. USA* **93**, 1913–1917 (1996).
6. Gelles, J., Schnapp, B. J. & Sheetz, M. P. Tracking kinesin-driven movements with nanometre-scale precision. *Nature* **331**, 450–453 (1988).
7. Beese, L., Stubbs, G. & Cohen, C. Microtubule structure at 18 Å resolution. *J. Mol. Biol.* **194**, 257–264 (1987).
8. Stewart, R. J., Thaler, J. P. & Goldstein, L. S. B. Direction of microtubule movement is an intrinsic property of the motor domains of kinesin heavy chain and *Drosophila* ncd protein. *Proc. Natl Acad. Sci. USA* **90**, 5209–5213 (1993).
9. Young, E. C., Berliner, E., Mahtani, H. K., Perez-Ramirez, B. & Gelles, J. Subunit interactions in dimeric kinesin heavy chain derivatives that lack the kinesin rod. *J. Biol. Chem.* **270**, 3926–3931 (1995).
10. Scholze, J. M., Heuser, J., Yang, J. T. & Goldstein, L. S. B. Identification of globular mechanochemical heads of kinesin. *Nature* **338**, 355–357 (1989).
11. Berliner, E., Young, E. C., Anderson, K., Mahtani, H. K. & Gelles, J. Failure of a single-headed kinesin to track parallel to microtubule protofilaments. *Nature* **373**, 718–721 (1995).
12. Berliner, E. thesis, Brandeis Univ. Waltham, MA (1995).
13. Howard, J., Hudspeth, A. J. & Vale, R. D. Movement of microtubules by single kinesin molecules. *Nature* **342**, 154–158 (1989).
14. Block, S. M., Goldstein, L. S. B. & Schnapp, B. J. Bead movement by single kinesin molecules studied with optical tweezers. *Nature* **348**, 348–352 (1990).
15. Vale, R. D. et al. Direct observation of single kinesin molecules moving along microtubules. *Nature* **380**, 451–453 (1996).
16. Segel, I. H. *Enzyme Kinetics* (Wiley, New York, 1975).
17. Gilbert, S. P., Webb, M. R., Brune, M. & Johnson, K. A. Pathway of processive ATP hydrolysis by kinesin. *Nature* **373**, 671–676 (1995).
18. Ma, Y.-Z. & Taylor, E. W. Mechanism of microtubule kinesin ATPase. *Biochemistry* **34**, 13242–13251 (1995).
19. Arnal, I., Metz, F., DeBonis, S. & Wade, R. Three-dimensional structure of functional motor proteins on microtubules. *Curr. Biol.* **6**, 1265–1270 (1996).
20. Hirose, K., Lockhart, A., Cross, R. A. & Amos, L. A. Three-dimensional cryoelectron microscopy of dimeric kinesin and ncd motor domains on microtubules. *Proc. Natl Acad. Sci. USA* **93**, 9539–9544 (1996).
21. Hirose, K., Lockhart, A., Cross, R. A. & Amos, L. A. Nucleotide-dependent angular change in kinesin motor domain bound to tubulin. *Nature* **376**, 277–279 (1995).
22. Kikkawa, M., Ishikawa, T., Wakabayashi, T. & Hirokawa, N. Three-dimensional structure of the kinesin head–microtubule complex. *Nature* **376**, 274–276 (1995).
23. Ray, S., Meyhoefer, E., Milligan, R. A. & Howard, J. Kinesin follows the microtubule's protofilament axis. *J. Cell Biol.* **121**, 1083–1093 (1993).
24. Gelles, J. et al. Structural and functional features of one- and two-headed biotinylated kinesin derivatives. *Biophys. J.* **68**, 276s–282s (1995).
25. Berliner, E. et al. Microtubule movement by a biotinylated kinesin bound to a streptavidin-coated surface. *J. Biol. Chem.* **269**, 8610–8615 (1994).
26. Hackney, D. D., Malik, A.-S. & Wright, K. W. Nucleotide-free kinesin hydrolyzes ATP with burst kinetics. *J. Biol. Chem.* **264**, 15943–15948 (1989).
27. Hackney, D. D. Kinesin ATPase: rate-limiting ADP release. *Proc. Natl Acad. Sci. USA* **85**, 6314–6318 (1988).
28. Gallagher, N. C. Jr & Wise, G. L. A theoretical analysis of the properties of median filters. *IEEE Trans. Acoust. Speech Signal Process.* **29**, 1136–1141 (1981).
29. Savitzky, A. & Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639 (1964).

Acknowledgements. We thank X. Liu for preliminary studies, H. Mahtani for technical assistance, and C. Miller and H. Huxley for comments on the manuscript. This work was supported by the NIH and by an HHMI predoctoral fellowship to E.C.Y.

Correspondence and requests for materials should be addressed to J.G. (e-mail: gelles@binah.cc.brandeis.edu).

A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity

Ruslan Medzhitov*, Paula Preston-Hurlburt & Charles A. Janeway Jr*

Section of Immunobiology, Yale University School of Medicine, and * Howard Hughes Medical Institute, New Haven, Connecticut 06520-8011, USA

Induction of the adaptive immune response depends on the expression of co-stimulatory molecules and cytokines by antigen-presenting cells. The mechanisms that control the initial induction of these signals upon infection are poorly understood. It has been proposed that their expression is controlled by the non-clonal, or innate, component of immunity that preceded in evolution the development of an adaptive immune system in vertebrates¹. We report here the cloning and characterization of a human homologue of the *Drosophila* toll protein (Toll) which has been shown to induce the innate immune response in adult *Drosophila*²⁻⁴. Like *Drosophila* Toll, human Toll is a type I transmembrane protein with an extracellular domain consisting of a leucine-rich repeat (LRR) domain, and a cytoplasmic domain homologous to the cytoplasmic domain of the human interleukin (IL)-1 receptor. Both *Drosophila* Toll and the IL-1 receptor are known to signal through the NF- κ B pathway⁵⁻⁷. We show that a constitutively active mutant of human Toll transfected into human cell lines can induce the activation of NF- κ B and the expression of NF- κ B-controlled genes for the inflammatory cytokines IL-1, IL-6 and IL-8, as well as the expression of the co-stimulatory molecule B7.1, which is required for the activation of naive T cells.

The Toll protein controls dorsal-ventral patterning in *Drosophila* embryos and activates the transcription factor Dorsal upon binding to its ligand Spatzle⁸. In adult *Drosophila*, the Toll/Dorsal signalling pathway participates in an anti-fungal immune response². Signalling through Toll parallels the signalling pathway induced by the IL-1 receptor (IL-1R) in mammalian cells: IL-1R signals through the NF- κ B pathway, and Dorsal and its inhibitor Cactus are homologous to NF- κ B and I- κ B proteins, respectively^{5,6}. Moreover, the cytoplasmic domain of *Drosophila* Toll is homologous to the cytoplasmic domain of IL-1R (ref. 9). Remarkably, the tobacco-virus-resistance gene that encodes N-protein is also similar to Toll in that it contains both a Toll signalling domain and an LRR domain¹⁰. It thus appears that the immune-response system mediated by Toll represents an ancient host defence mechanism⁶ (Fig. 1). To investigate the possibility that this pathway has been retained in the immune system of vertebrates, we used sequence and pattern searches¹¹ of the expressed-sequence tag (EST) database at the National Center for Biotechnology Information (NCBI). A search with a sequence profile of the Toll/IL-1R signalling domain identified a matching sequence in the EST database derived from human fetal liver/spleen library (Genebank accession number H48602, corresponding to clone 202057 from the IMAGE consortium¹²). Sequencing of this clone showed that it corresponds to the 3' untranslated region (UTR) and the C terminus of the coding region of the messenger RNA. A fragment of the clone amplified by using the polymerase chain reaction (PCR) was used to screen a human spleen complementary DNA library by hybridization, and the 5' end of the cDNA was cloned using the 5'-RACE technique as described^{13,14}. The full-length 4,874-base-pair (bp) cDNA clone contained an open reading frame of 2,523 bp which encoded an

841-amino-acid protein chain (Fig. 2a), as well as a LINE-1 reverse transcriptase pseudogene in the 3'-UTR. Alignment of the sequences of the human and *Drosophila* Toll proteins shows that there is homology over the entire length of the protein chains (Fig. 2b). Notably, the similarity between the cytoplasmic domains is higher than between the human proteins Toll and IL-1R (not shown). The extracellular domain of human Toll (hToll) contains 21 tandemly repeated leucine-rich motifs separated by a non LRR region, similar to *Drosophila* Toll (dToll). At the N-terminal end of the LRR domain there is a 31-amino-acid long N-flanking region that is also present in several other LRR-containing proteins, for example RP105, Decorin and Biglycan¹⁵. The C-terminal end of the LRR domain is flanked by a cysteine-rich domain which is also present in dToll and some other transmembrane proteins^{16,17}.

To examine the expression pattern of hToll, a 720-bp cDNA fragment was used to probe northern blots containing poly(A)⁺ RNA from several organs. Most organs expressed two mRNA species: one of ~5 kilobases (kb) was predominant in most tissues except peripheral blood leukocytes (PBL), and corresponded to the length of the cDNA that we cloned. The lower band was ~4 kb long and this band was predominant in the PBL. The 4-kb band was not detectable in kidney, and liver did not contain any mRNA at all (Fig. 3). We also tested different mouse and human cell lines for expression of hToll mRNA by using PCR with reverse transcription (RT-PCR). We found mRNA for hToll in monocytes, macrophages, dendritic cells, γ/δ T cells, Th1 and Th2 α/β T cells, a small intestinal epithelial cell line, and a B-cell line (data not shown). The hToll gene is expressed most strongly in spleen and PBL (Fig. 3); its expression in other tissues may be due to the presence of macrophages and dendritic cells, in which it could act as an early-warning system for infection. Alternatively, hToll may be widely expressed because hToll signals through the conserved NF- κ B pathway (see below) and NF- κ B is a ubiquitous transcription factor.

To characterize hToll functions and see whether it can induce transcription of immune response genes like dToll, we generated a dominant-positive mutant of hToll because the natural ligand of hToll is unknown. To produce a constitutively active mutant of hToll, we made use of genetic information from dToll: analysis of ventralizing mutants in *Drosophila* embryos had identified the function of the ectodomain C-flanking cysteine-rich region in dToll¹⁶ as controlling the activity of dToll in signal transduction. In three dominant

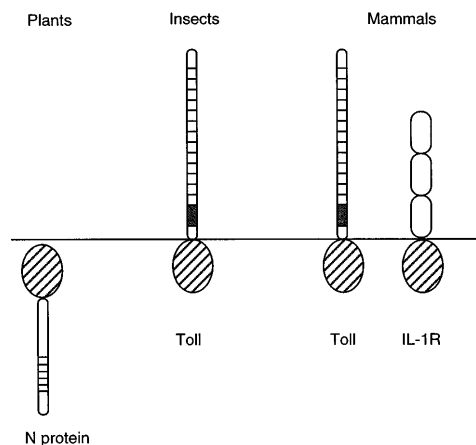


Figure 1 Ancient immune defence systems of plants, insects and vertebrates. A homologous immune response system based on the Toll signalling domain is used in plants, insects and vertebrates. In mammals, Toll induces signals required for the activation of both an innate and an adaptive immune response (see text). The figure is modified from ref. 6. Diagonal hatching represents the Toll signalling domain; striped rods, leucine-rich repeat domains; black rectangles, C-terminal cysteine-rich domains; white ellipsoids, immunoglobulin domains.

mutations, one of four conserved cysteine residues was changed to a tyrosine, rendering dToll constitutively active¹⁶. In another dominant mutant allele of dToll, this region is deleted altogether, together with the rest of the extracellular part of dToll¹⁸, indicating that a constitutively active dominant mutant of dToll can be generated by either disrupting the disulphide bonds in the C-terminal region flanking the ectodomain or deleting this region and the LRR region of the ectodomain^{16,18}. As this region is conserved in hToll (Fig. 2b), we constructed a recombinant hToll receptor that mimics features of the dominant-positive mutants of dToll in which the ectodomain of the mouse CD4 (mCD4) molecule was spliced with the transmembrane and cytoplasmic domains of hToll so that almost all of the ectodomain of hToll (Fig. 4a) and most of the C-terminal flanking region, including three of the four conserved cysteine residues, were removed. Mouse CD4 acted as an epitope tag to monitor the expression of the construct in human cells by using monoclonal antibodies. This chimaeric construct was cloned into an expression vector and transfected into the human monocytic cell line THP-I which endogenously expresses hToll and should thus contain the downstream signalling components. Seven stably transfected clones were selected for further analysis. Expression of the CD4/hToll chimaera in these clones was identified by surface mCD4 expression detected by FACS analysis (Fig. 4b). We used RT-

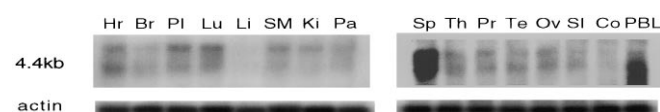
PCR on seven independent clones to test whether the dominant-positive mutant of hToll can induce expression of NF- κ B-controlled immune-response genes; untransfected or mock-transfected THP-I cells acted as controls. Each experiment was repeated 8–12 times and the results from two representative CD4/hToll transfected clones and untransfected THP-I cells are shown in Fig. 4c. Expression of a constitutively active mutant of hToll caused induction of the IL-1, IL-8 and B7.1 genes. In the presence of interferon (IFN)- γ , IL-6 mRNA was also strongly induced, which was not due to induction of IL-1 by IFN- γ because during the time of the experiments (8–20 hours) IL-1 was induced in control cells but IL-6 was not. Other genes that might be controlled by signalling through hToll, for example inducible nitric oxide synthase (iNOS) and B7.2, although expressed in the transfectants, could not be tested in this system because of constitutive expression (B7.2) or clonal heterogeneity in the bulk population of control cells (iNOS).

As dToll signals through the NF- κ B pathway, we tested whether hToll could induce activation of NF- κ B. The CD4/hToll chimaeric construct was transiently transfected into Jurkat cells together with an NF- κ B-promoter-driven luciferase reporter gene. As shown in Fig. 4d, expression of CD4/hToll causes induction of NF- κ B, confirming that Toll/NF- κ B signalling is conserved in *Drosophila* and humans.



Figure 2 a, Amino-acid sequence of human Toll. The predicted signal peptide (residues 1–23) is underlined. The predicted transmembrane segment (residues 634–653) is in bold and underlined. The cDNA sequence (Genbank accession U93091) contains two coding regions: one has an open reading frame (position 104–2,627) and encodes the 841-amino-acid human Toll. The second coding region (position 3,970–4,489) contains a pseudogene, *LINE-1* reverse-transcriptase homologue (not shown). **b**, Alignment of the protein sequences of hToll (top) and dToll (bottom). A vertical line indicates identity and a colon represents similarity between the corresponding pair of amino-acid residues. The location of the cysteine-rich domain is indicated by underlining in the *Drosophila* sequence, and the site used to prepare chimaeric hToll is indicated by an asterisk above the hToll sequence. Parameters of the alignment were as follows: Ktuple: 2, gap penalty: 3, gap length penalty: 8.

Figure 3 Northern blots of poly(A)⁺ RNA demonstrate predominant expression of hToll in spleen (Sp) and peripheral blood leukocytes (PBL). The upper band corresponds to the length of the hToll cDNA we have cloned (~5 kb); the lower band is ~4 kb. The difference in the sizes of the two bands is probably due to the differential processing of the 3'UTR, as detected by RT-PCR analysis of this region (not shown) and might be due to the presence of the *LINE-1* reverse-transcriptase pseudogene in this region. Tissues examined: Hr: heart; Br: brain; Pl: placenta; Lu: lung; Li: liver; SM: skeletal muscle; Ki: kidney; Pa: pancreas; Th: thymus; Pr: prostate; Te: testis; Ov: ovary; SI: small intestine; Co: colon.



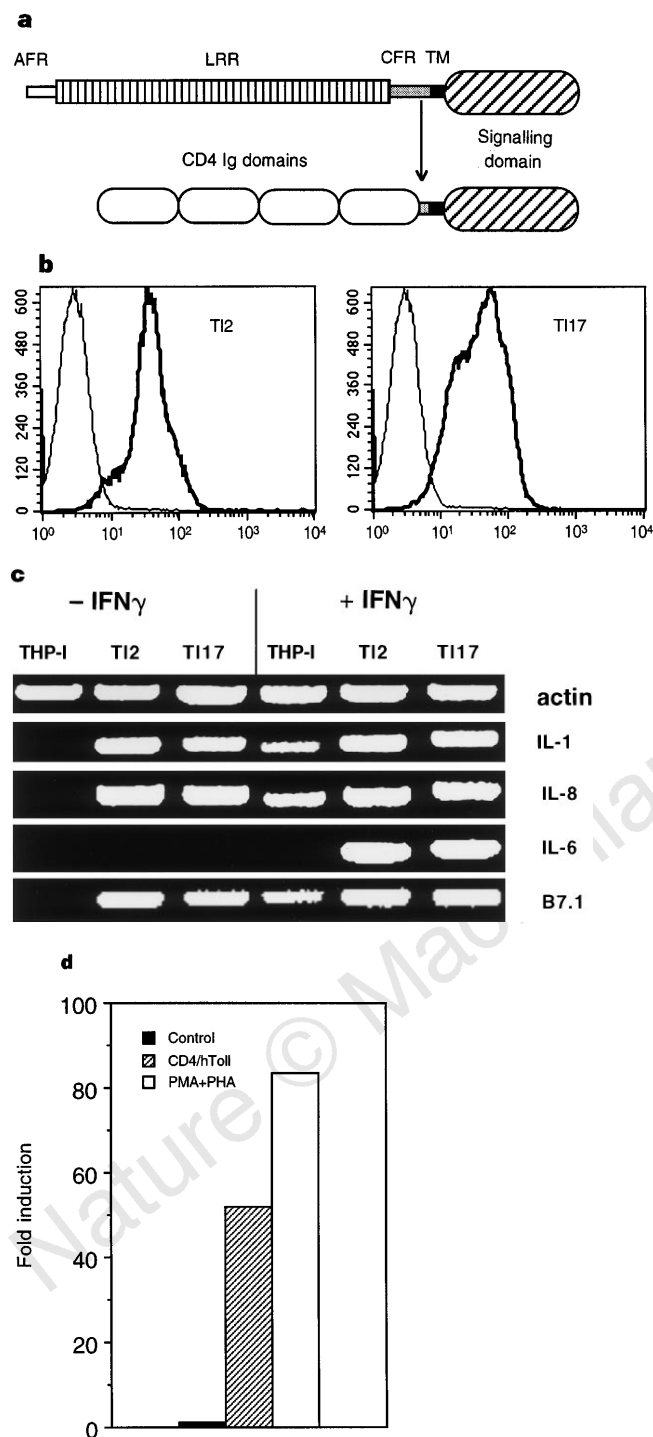


Figure 4 Signalling function of hToll. **a**, Representation of the mouse CD4/hToll chimera that renders the signalling domain of hToll constitutively active. hToll cut at the site of the asterisk in Fig. 2b. AFR and CFR represent the amino-terminal flanking and carboxy-terminal cysteine-rich regions. TM, transmembrane region. **b**, FACS analysis of mCD4/hToll transfected clones TI2 (left) and TI17 (right, dark lines) stained with a monoclonal antibody for mouse CD4. Untransfected THP-1 cells (fine line) stained with the same antibody were used as a control. **c**, RT-PCR analysis of the gene expression induced by signalling through hToll. IFN- γ treatment was 8–20 h in various experiments which gave the same results using 20 ng human recombinant IFN- γ (Gibco BRL). **d**, Activation of NF- κ B by the dominantly active CD4/hToll chimera in Jurkat cells, as measured by reporter gene activity. Controls were transfected with empty vector (negative controls) and activated with phorbol myristyl acetate (PMA; 20 ng ml⁻¹) and phytohaemagglutinin (PHA; 2 μ g ml⁻¹) for 3–4 h. All transfectants contained a pB2XLuc reporter construct.

These findings indicate that Toll functions in vertebrates as a non-clonal receptor of the immune system. Pro-inflammatory cytokines such as IL-1, IL-6 and IL-8 are induced early during infection but, apart from CD14, the cellular receptors that are activated during this initial phase have not yet been identified. It has been argued that the innate immune response has been preserved in vertebrates not only to resist infection before adaptive immunity is induced, but also to induce signals that inform the adaptive system of the presence of pathogens within the internal milieu^{1,19,20}. We have shown that the Toll/NF- κ B host defence pathway is conserved from *Drosophila* to humans and that the human homologue of the *Drosophila* Toll protein can induce signals for activating both an innate and an adaptive immune response in vertebrates. □

Methods

Cloning and sequence analysis. BLAST searches of the NCBI sequence database were made using sequence profiles of the dToll/IL-1R signalling domains. Matching sequence (Genbank accession number H48602) corresponded to the IMAGE consortium clone number 202057. PCR primers were designed based on the EST sequence and used to amplify a 220-bp fragment from a human spleen cDNA library (Clontech). The PCR-amplified fragment was used to screen a human spleen cDNA library as described¹³. The 5' end of the cDNA was cloned by 5'-RACE protocol as described¹⁴. Cloned products were subjected to automated sequencing (ABI).

RNA expression. Tissue northern blots were purchased from Clontech and analysed according to the manufacturer's instructions. The presence of equal amounts of RNA on the blots was confirmed by hybridization with a β -actin probe.

RT-PCR analysis. Total RNA was isolated using TRIzol Reagent (Gibco BRL) and 0.5 μ g was used for reverse transcription using AMV reverse transcriptase (Seikagaku America) or Superscript reverse transcriptase (Gibco BRL). One-tenth to one-twentieth of this reaction was used as a template for PCR amplification with *Taq* polymerase (Gibco) for 25–35 cycles at 94 °C for 1 min, 55 °C for 1 min, and 72 °C for 1 min. Sequences of the primers were taken from ref. 21.

Construction of mCD4/hToll chimera. An *Sph*I site was introduced between V365 and N366 in the mouse CD4 gene by site-directed mutagenesis. An internal *Sph*I site in hToll at the position between M620 and P621 was used for the construction of the chimera.

Monoclonal antibodies and FACS analysis. Anti-mouse CD4 antibody (L3T4) (Pharmingen) was used for staining. Flow cytometry analysis was done on a FACScan (Becton–Dickinson).

NF- κ B assay. Jurkat cells were transiently transfected using Lipofectamine reagent (Gibco BRL) according to the manufacturer's instructions. 1 μ g pB2XLuc plasmid, containing NF- κ B-driven luciferase gene was cotransfected with 1 μ g pSR α N expression vector with or without (control sample) the insert encoding CD4/Toll chimera. For a positive control, cells were treated with PMA (phorbol myristyl acetate; 20 ng ml⁻¹) and PHA (phytohaemagglutinin; 2 μ g ml⁻¹) for 3–4 h. Cells were lysed 2 or 3 days later for measurement of luciferase activity using reagents from Promega.

Received 17 March; accepted 21 May 1997.

- Janeway, C. A. Jr Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harbor Symp. Quant. Biol.* **54**, 1–13 (1989).
- Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M. & Hoffmann, J. A. The dorsoventral regulatory gene cassette *spatzle/toll/cactus* controls the potent antifungal response in *Drosophila* adults. *Cell* **86**, 973–983 (1996).
- Rosetto, M., Engström, Y., Baldari, C. T., Telford, J. L. & Hultmark, D. Signals from the IL-1 receptor homolog, Toll, can activate an immune response in a *Drosophila* hemocyte cell line. *Biochem. Biophys. Res. Commun.* **209**, 111–116 (1995).
- Ip, Y. T. & Levine, M. Molecular genetics of *Drosophila* immunity. *Curr. Opin. Genet. Dev.* **4**, 672–677 (1994).
- Wasserman, S. A. A conserved signal transduction pathway regulating the activity of rel-like proteins dorsal and NF- κ B. *Mol. Biol. Cell* **4**, 767–771 (1993).
- Hultmark, D. Insect immunology: Ancient relationships. *Nature* **367**, 116–117 (1994).
- Belvin, M. P. & Anderson, K. V. A conserved signaling pathway: the *Drosophila* Toll–dorsal pathway. *Annu. Rev. Cell Dev. Biol.* **12**, 393–416 (1996).
- Morisato, D. & Anderson, K. V. The *spatzle* gene encodes a component of the extracellular signaling pathway establishing the dorsal–ventral pattern of the *Drosophila* embryo. *Cell* **76**, 677–688 (1994).
- Gay, N. J. & Keith, F. J. *Drosophila* Toll and IL-1 receptor. *Nature* **351**, 355–356 (1991).
- Whitham, S. et al. The product of the tobacco mosaic virus resistance gene N: similarity to toll and the interleukin-1 receptor. *Cell* **78**, 1101–1115 (1994).
- Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119–129 (1994).

12. Lennon, G., Auffray, C., Polymeropoulos, M. & Soares, M. B. The IMAGE Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33**, 151–152 (1996).
13. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, New York, 1989).
14. Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. (eds) *PCR Protocols: A Guide to Methods and Applications* (Academic, Orlando, 1992).
15. Miyake, K., Yamashita, Y., Ogata, M., Sudo, T. & Kimoto, M. RP105, a novel B cell surface molecule implicated in B cell activation, is a member of the leucine-rich repeat protein family. *J. Immunol.* **154**, 3333–3340 (1995).
16. Schneider, D. S., Hudson, K. L., Lin, T. Y. & Anderson, K. V. Dominant and recessive mutations define functional domains of Toll, a transmembrane protein required for dorsal–ventral polarity in the *Drosophila* embryo. *Genes Dev.* **5**, 797–807 (1991).
17. Takahashi, N., Takahashi, Y. & Putnam, F. W. Periodicity of leucine and tandem repetition of a 24-amino acid segment in the primary structure of leucine-rich alpha 2-glycoprotein of human serum. *Proc. Natl Acad. Sci. USA* **82**, 1906–1910 (1985).
18. Winans, K. A. & Hashimoto, C. Ventralization of the *Drosophila* embryo by deletion of extracellular leucine-rich repeats in the Toll protein. *Mol. Biol. Cell.* **6**, 587–596 (1995).
19. Fearon, D. T. & Locksley, R. M. The instructive role of innate immunity in the acquired immune response. *Science* **272**, 50–53 (1996).
20. Medzhitov, R. & Janeway, C. A. Jr On the semantics of immune recognition. *Res. Immunol.* **147**, 208–214 (1996).
21. Yamamura, M. *et al.* Defining protective responses to pathogens: cytokine profiles in leprosy lesions. *Science* **254**, 277–279 (1991).

Acknowledgements. We thank our colleagues for discussions; P. G. Waterbury for DNA sequencing; S. Ghosh and R. Voll for pB2XLuc plasmid and for helpful suggestions; and J. Flaxenberg for assistance. We acknowledge the Howard Hughes Medical Institute and the NIAID, NIH for support of this work.

Correspondence and requests for materials should be addressed to C.A.J.Jr (e-mail: charles.janeway@yale.edu). The sequence is deposited in Genebank data base under the accession number U93091.

A protein related to splicing factor U2AF³⁵ that interacts with U2AF⁶⁵ and SR proteins in splicing of pre-mRNA

Hélène Tronchère*, Jiwu Wang* & Xiang-Dong Fu

Division of Cellular and Molecular Medicine, Department and School of Medicine, University of California at San Diego, 9500 Gilman Drive, La Jolla, California 92093-0651, USA

* These authors contributed equally to this work.

Recognition of a functional 3' splice site in pre-mRNA splicing requires a heterodimer of the proteins U2AF⁶⁵/U2AF³⁵. U2AF⁶⁵ binds to RNA at the polypyrimidine tract^{1,2}, whereas U2AF³⁵ is thought to interact through its arginine/serine-rich (RS) domain with other RS-domain-containing factors bound at the 5' splice site, assembled in splicing enhancer complexes, or associated with the U4/U6.U5 small nuclear ribonucleoprotein complex^{3–7}. It is unclear, however, how such network interactions can all be established through the small RS domain in U2AF³⁵. Here we describe the function of a U2AF³⁵-related protein (Urp), which is the human homologue of a mouse imprinted gene. Nuclear extracts depleted of Urp are defective in splicing, but activity can be restored by addition of recombinant Urp. U2AF³⁵ could not replace Urp in complementation, indicating that their functions do not overlap. Co-immunodepletion showed that Urp is associated with the U2AF⁶⁵/U2AF³⁵ heterodimer. Binding studies revealed that Urp specifically interacts with U2AF⁶⁵ through a U2AF³⁵-homologous region and with SR proteins (a large family of RS-domain-containing proteins) through its RS domain. Therefore, Urp and U2AF³⁵ may independently position RS-domain-containing factors within spliceosomes.

We previously purified and cloned a mammalian serine kinase, SRPK1, which is highly specific for the RS-rich superfamily of splicing factors^{8–10}. This stringent substrate specificity enabled us to do a yeast two-hybrid screen to search for new RS-domain-containing proteins. We isolated a positive clone from a HeLa complementary DNA library that contains a typical RS domain at the carboxy terminus (Fig. 1a). The recombinant protein was efficiently phosphorylated in the RS domain, like other RS-domain proteins, by

SRPK1 *in vitro* (data not shown). This protein of 483 amino acids does not contain any RNA-recognition motif characteristic of the SR family of splicing factors^{11,12}; instead, it contains two stretches of sequences, H1 and H2 (boxed in Fig. 1b), homologous to the human splicing factor U2AF³⁵. We therefore named it Urp, for U2AF³⁵-related protein. Northern blotting analysis indicates that both U2AF³⁵ and Urp are ubiquitously expressed in all human tissues and cell lines examined (data not shown). Homology searches revealed that Urp had been independently cloned from humans as U2AF1-RS2 (ref. 13), which is nearly identical (94% amino-acid identity) to another human gene, U2AF1-RS1 (ref. 13) (Fig. 1b). These two human genes are equally homologous (~80% identical) to two mouse genes¹⁴, one of which is imprinted¹⁵, and the other has a high transmission distortion in interspecific backcross progeny¹⁶. Thus, several U2AF³⁵-related proteins are expressed in mammalian cells and may be members of the superfamily of RS-domain-containing splicing factors¹¹.

To determine the function of Urp, we raised rabbit polyclonal antibodies against recombinant Urp expressed in bacteria. Immunoblotting of HeLa cell nuclear extracts demonstrated that anti-Urp antibodies specifically detect a single band of relative molecular mass 66K (Fig. 2a), which corresponds to endogenous Urp as it matches the size of Urp translated in reticulocyte lysate (data not shown). To see whether Urp is essential for splicing, we constructed affinity columns with anti-Urp antibodies or preimmune serum. Control depletion with preimmune serum had no effect, but passing nuclear extract through an anti-Urp affinity column efficiently depleted it of Urp, as evidenced by western blotting and complete inhibition of splicing (data not shown). Although this result is consistent with Urp being a splicing factor, this Urp-depleted extract could not be complemented by recombinant Urp, probably because of co-depletion of other essential splicing factors (see below). We therefore did the immunodepletion in the presence of high salt (0.5 M KCl) to minimize co-depletion. As shown in Fig. 2c, human β -globin pre-mRNA splicing in Urp-depleted nuclear extract was significantly decreased (compare lanes 1 and 2). The second step of the splicing reaction, which generates released lariat intermediate and

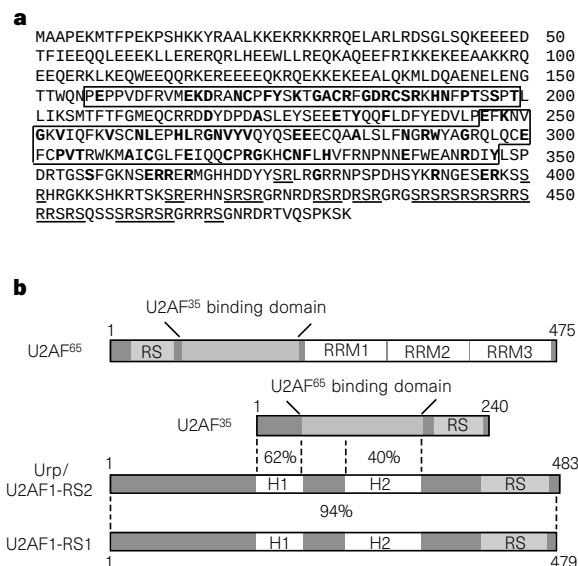


Figure 1 Urp structure and sequence comparison with U2AF³⁵. **a**, Amino-acid sequence of Urp. It is identical to U2AF1-RS2 (GenBank D49677; ref. 13). The two U2AF³⁵ homologous domains, H1 and H2, are boxed and amino-acid residues identical to those in U2AF³⁵ are shown in bold. Repeats of SR or RS dipeptides are underlined. **b**, Domain structure of the large and small subunits of U2AF and comparison of human U2AF³⁵, Urp/U2AF1-RS2 and U2AF1-RS1. Percentage identity between homologous domains is indicated.

ligated exons, was preferentially affected by Urp depletion under these conditions, indicating that Urp may play a role in this step; however, we could not rule out the possibility that it was required for the first step as we did not detect any accumulation of first-step splicing intermediates. To confirm the essential function of Urp in splicing, we restored Urp activity to a Urp-depleted nuclear extract with a glutathione-S-transferase (GST)-Urp fusion protein produced in baculovirus. Because Urp was insoluble in baculovirus-infected cells, we purified recombinant Urp from inclusion bodies, then isolated the protein by SDS-PAGE (Fig. 2b). After denaturation and renaturation in the presence of bovine serum albumin, recombinant Urp restored splicing activity to the Urp-depleted nuclear extract in a concentration-dependent manner (Fig. 2c, lanes 4 to 6). There was no complementing activity associated with BSA (Fig. 2c, lane 3). We conclude that Urp is an essential splicing factor.

SR proteins have redundant functions, at least *in vitro* (for review, see refs 11, 12). U2AF³⁵ was originally shown to be non-essential for constitutive splicing because U2AF⁶⁵ alone was sufficient to reconstitute a splicing-deficient nuclear extract from which the U2AF heterodimer was removed by poly(U) in the presence of high salt¹⁷. However, U2AF³⁵ is essential for constitutive and enhancer-dependent splicing by immunodepletion and reconstitution experiments⁵. Although the *Drosophila* homologue of U2AF³⁵ is required for viability¹⁸, there may be factors that are partially redundant with U2AF³⁵. Urp is such a candidate because it is related to U2AF³⁵ in structure and sequence. To test this possibility, we expressed and purified U2AF³⁵ in the same way as Urp (Fig. 2b), and found that U2AF³⁵ could not complement the Urp-depleted nuclear extract (Fig. 2c, lanes 7 to 9), indicating that Urp and U2AF³⁵ are not

functionally redundant. Results were identical when the complementation was repeated with a U2AF³⁵ protein preparation that had previously been shown to be functional in restoring splicing activity to a U2AF³⁵-depleted extract (ref. 5, and data not shown). Surprisingly, splicing in the Urp-depleted nuclear extract could be stimulated reproducibly by U2AF⁶⁵ (Fig. 2c, lanes 10 to 12). It has been found that U2AF⁶⁵ can stimulate splicing in a U2AF³⁵-deficient nuclear extract, probably as a result of recruiting the remaining U2AF³⁵ in the depleted extract⁵. U2AF⁶⁵ may therefore also interact with Urp, recruiting the remaining Urp in the Urp-depleted extract.

To test this, we examined the abundance of various splicing factors in the Urp-depleted nuclear extract by western blotting. As shown in Fig. 3a, Urp was effectively depleted by the anti-Urp affinity column in comparison to Urp present in untreated as well as preimmune serum-treated nuclear extracts. A fraction of U2AF⁶⁵ was co-depleted, consistent with an interaction between U2AF⁶⁵ and Urp. In contrast, the level of U2AF³⁵ was only slightly reduced, indicating that a large fraction of U2AF³⁵ is not associated with Urp or complexed with U2AF⁶⁵, in agreement with previous functional studies on U2AF³⁵ (ref. 5). The amount of SR proteins detected by monoclonal antibody 104 or of snRNP-associated proteins detected by anti-Sm antibodies was not affected by Urp depletion. The co-immunodepletion of Urp and U2AF⁶⁵ indicates that Urp forms a functional complex with U2AF⁶⁵ *in vivo*. This conclusion does not contradict the observation that U2AF⁶⁵ purifies only as a heterodimer with U2AF³⁵, because 2M guanidine-HCl was used in a

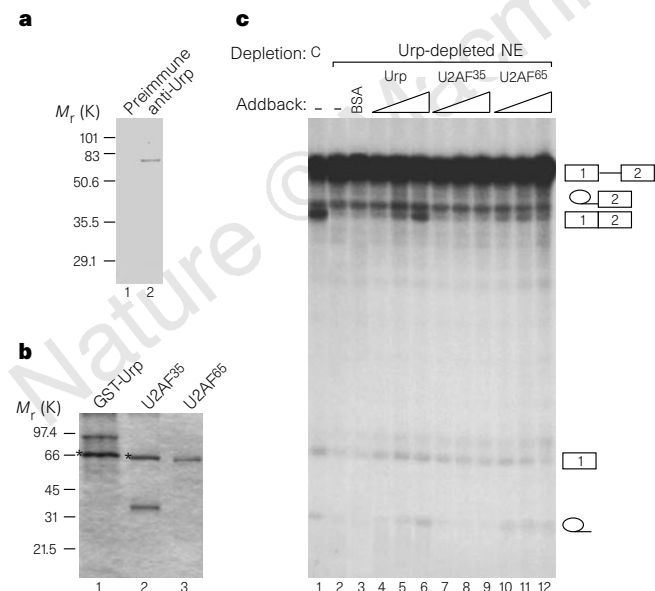


Figure 2 Urp is an essential splicing factor. **a**, Urp was detected in HeLa nuclear extracts (25 µg protein per lane) by anti-Urp antibodies (1:3,000) but not by the preimmune serum (1:1,000). **b**, Recombinant proteins used for complementation. All proteins were expressed in Sf9 cells. GST-Urp (lane 1) and His-U2AF³⁵ (lane 2) were purified by SDS-PAGE. Specific protein concentration was adjusted to 0.1 mg ml⁻¹. BSA (asterisks) was used as a carrier during gel elution and denaturation and renaturation. His-U2AF⁶⁵ (lane 3) was directly purified from sonicated cell lysate on a Ni²⁺ column (Invitrogen). **c**, Splicing of human β-globin pre-mRNA in control depleted (lane 1), or Urp-depleted (lanes 2-12) HeLa cell nuclear extracts (NE). The Urp-depleted nuclear extract was complemented with BSA control (9 µl, lane 2), or with three concentrations (1, 3, 9 µl) of GST-Urp (lanes 4-6), His-U2AF³⁵ (lane 7-9), or His-U2AF⁶⁵ (lane 10-12). On the right are the structures of human β-globin splicing substrate, intermediates and products: boxes, exon 1 and exon 2; thin line, intron.

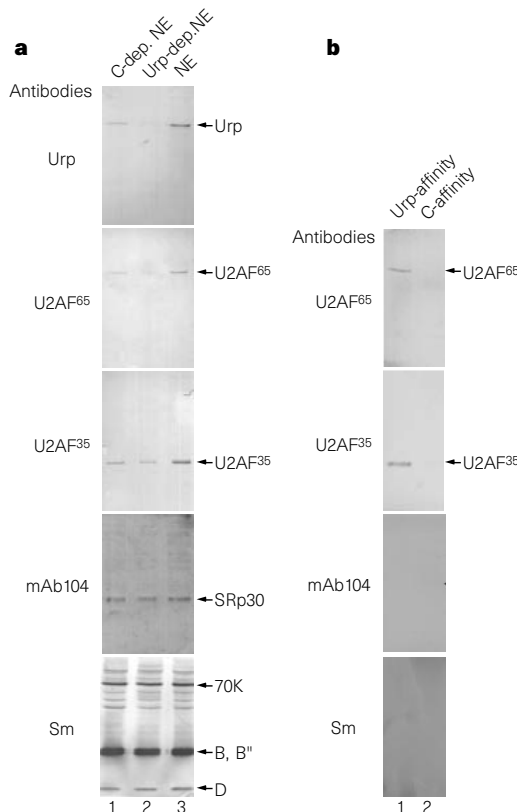


Figure 3 Urp is associated with the U2AF heterodimer. **a**, 2 µl of nuclear extracts that had been depleted with preimmune serum (lane 1), anti-Urp antibodies (lane 2), or untreated (lane 3) were probed by western blotting with anti-Urp antibodies (1:3,000), anti-U2AF⁶⁵ (1:5,000), anti-U2AF³⁵ (1:5,000), mAb104 (culture supernatant), or anti-Sm (Y12, culture supernatant), as indicated on the left. Specific proteins detected by these antibodies are indicated on the right. **b**, Anti-Urp and control affinity columns (0.5 ml bed volume) were eluted with 0.5 ml of PBS plus 2M NaCl. 10 µl of eluant from each column was probed with the antibodies indicated on the left; and proteins detected are indicated on the right.

purification step¹⁹: these extreme conditions were not used in Urp depletion, so a weaker association between Urp and U2AF⁶⁵ was likely to be preserved.

We also tested for any association of Urp with other splicing factors by western blotting analysis of high-salt eluants from the control and anti-Urp affinity columns. As expected, U2AF⁶⁵, but not SR proteins or Sm antigens, was specifically detected in the eluant of the anti-Urp affinity column (Fig. 3b). U2AF³⁵ seemed to be stoichiometrically present in the Urp–U2AF⁶⁵ complex, because similar amounts of U2AF⁶⁵ and U2AF³⁵ were detected by specific antibodies of similar titres. This can be explained by the specific association of Urp with the U2AF heterodimer. Alternatively, Urp may be independently associated with equal amounts of U2AF⁶⁵ and U2AF³⁵, but we consider this less likely as only U2AF⁶⁵ binds strongly to Urp *in vitro* and in the yeast two-hybrid system (see below).

We next assayed the binding between Urp and U2AF to investigate the role of Urp in splicing. We expressed GST–U2AF³⁵ and GST–U2AF⁶⁵ in bacteria, and carried out a GST-pulldown assay using *in vitro* translated Urp, U2AF³⁵ and U2AF⁶⁵ (Fig. 4a). As expected, *in vitro* translated U2AF⁶⁵ bound strongly to GST–U2AF³⁵ immobilized on glutathione–Sephadex beads but not to vacant beads, demonstrating the specificity of binding. Under the same conditions (0.4M KCl), both Urp and U2AF³⁵ bound efficiently to immobilized U2AF⁶⁵ (Fig. 4a). The interaction between Urp and U2AF⁶⁵ was not mediated by RNA because the extent of binding was similar in the presence of RNase A (data not shown).

Interaction between Urp and U2AF⁶⁵ could be mediated by the RS domain present in both proteins as this domain is a known protein–protein–interaction interface^{3,4}, or by one or both of the U2AF³⁵ homologous domains (H1 and H2). The H2 domain in Urp is a strong candidate because the corresponding region in U2AF³⁵ is known to interact with U2AF⁶⁵ (ref. 20). We therefore prepared and tested a series of Urp mutants (Fig. 4b). Several conclusions can be drawn from the data shown in Fig. 4c and quantified in Fig. 4d. First, wild-type Urp binds to U2AF⁶⁵, but its binding is less efficient than that between U2AF⁶⁵ and U2AF³⁵; the Urp–U2AF⁶⁵ interaction was prevented in the presence of 1 M KCl (data not shown). This

explains why Urp is associated with U2AF⁶⁵ only under immuno-depletion conditions and not under the more stringent purification conditions¹⁹. Second, deletion of the RS domain from Urp has no effect on its binding to U2AF⁶⁵, indicating that the RS domain is not involved in the interaction between the two proteins. Third, although deletion of the H1 domain has no effect, deletion of the H2 domain significantly, but not completely, diminished Urp binding to U2AF⁶⁵. Thus, U2AF³⁵ and Urp seem to interact with U2AF⁶⁵ through homologous sequences. However, the interaction between Urp and U2AF⁶⁵ could not be competed by increasing amounts of unlabelled U2AF³⁵ (data not shown), supporting the idea that Urp interacts with the U2AF heterodimer *in vivo*.

To verify the sequence requirements for protein–protein interaction between Urp and U2AF⁶⁵, and to determine whether Urp, like U2AF³⁵, also interacts with SR proteins through its RS domain, we extended the binding studies to the yeast two-hybrid system. As shown in Fig. 4e, Urp interacts with U2AF⁶⁵ in yeast and deletion of the H2 domain diminishes binding, but deletion of the RS domain in Urp has no effect, consistent with the *in vitro* results. Urp, like U2AF³⁵, interacts with two SR proteins, SC35 and SF2/ASF, but not with the U1-70k protein (data not shown); deletion of the RS domain in Urp reduces its interaction with either SR protein, indicating that the interaction between Urp and SR proteins is mediated by their RS domains (Fig. 4e). Deletion of the H2 domain had no effect on the interaction between Urp and SC35. However, for unknown reasons, the removal of the H2 domains from Urp decreased its interaction with SF2/ASF. We could not detect any significant interaction between U2AF³⁵ and SF2/ASF, in contrast to a previous report³. Under the same conditions, Urp interacts with SF2/ASF, indicating a preference for SF2/ASF to bind with Urp rather than U2AF³⁵ (Fig. 4e).

We have shown that Urp is an essential splicing factor. On the basis of our finding that Urp is associated with the U2AF heterodimer in nuclear extracts and interacts with U2AF⁶⁵ and SR proteins through distinct domains, we propose that Urp may form part of a larger U2AF complex and be engaged in network interactions during spliceosome assembly. Because Urp seems to play a non-redundant function with

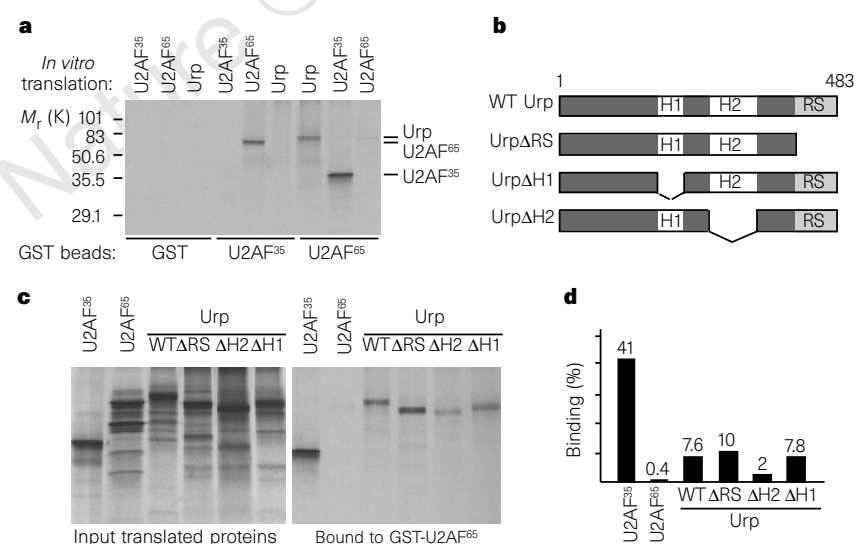


Figure 4 Urp interacts with U2AF⁶⁵ and SR proteins *in vitro* and in yeast. **a**, Identical amounts of *in vitro* translated ³⁵S-labelled Urp, U2AF⁶⁵ and U2AF³⁵ (top) were incubated with GST, GST–U2AF³⁵ or GST–U2AF⁶⁵ immobilized on glutathione beads (bottom). Bound proteins are indicated on the right. On the left are low-range prestained M_r markers (Bio-Rad). **b**, Organization of wild-type (WT) Urp and three deletion mutants. **c**, Binding of wild-type and mutant Urp to immobilized GST–U2AF⁶⁵. **d**, Quantification of binding efficiency. **e**, Pairwise protein–protein interaction in the yeast LexA two-hybrid system. Each number represents the average of data derived from at least three independent transformations and is expressed as fold activation above the background determined using specific bait and the empty prey vector.

Protein-protein interaction in the yeast two-hybrid system					
bait \ prey	U2AF ³⁵	Urp	UrpΔRS	UrpΔH2	U1-70K
U2AF ⁶⁵	68.0	43.0	33.6	2.4	1.1
SC35	6.5	12.8	2.4	13.7	24.1
ASF/SF2	1.0	9.2	0.7	3.0	17.5

U2AF³⁵ and appears preferentially to affect the second step of the splicing reaction, it may interact with a selective set of factors during reaction. The unique behaviour of Urp homologues in mouse indicates that Urp and its family members may be targets for regulation. □

Methods

Cloning, expression and purification. SRPK1 was cloned in pGBT9 as bait to screen a HeLa cDNA library (from G. Hannon) constructed in the pGAD-GH vector as described²¹. The initial Urp clone was lacking 92 amino acids from its N terminus, and full-length Urp cDNA was isolated from a λgt11 HeLa S3 cDNA library (Clontech).

Urp was expressed as GST fusion proteins either in bacteria or by baculovirus. Thrombin-cleaved bacterial Urp was used to raise rabbit antibodies, diluted 1:3,000 for western blotting. For splicing complementation, baculovirus-expressed Urp and U2AF were purified from the inclusion bodies of Sf9 cells, followed by isolation from SDS-PAGE. Briefly, 10⁸ infected Sf9 cells were sonicated in 10 ml buffer A (20 mM Tris, pH 7.6, 150 mM KCl, 2 mM EDTA, 1 mM DTT) plus standard protease inhibitors. Samples were centrifuged through a 15-ml sucrose cushion (40% sucrose in buffer A) at 12,000 r.p.m. for 30 min in a HB-4 rotor. Pellets were resuspended in 1 ml buffer A plus 8 M urea. An estimated 10 µg of specific protein was loaded per well onto a 10% SDS-polyacrylamide gel. Proteins were eluted from gel slices in buffer B (50 mM Tris, pH 7.5, 0.1% SDS, 0.1 mg ml⁻¹ BSA, 1 mM DTT, 0.2 mM EDTA, 0.1 mM PMSF and 2.5% glycerol) overnight at 4 °C. Eluted proteins were precipitated with four volumes of cold acetone overnight at -20 °C. After a 15-min spin, the pellet was washed with cold methanol, dried and resuspended in 2.5 µl 8 M urea. Protein was renatured by adding 125 µl buffer C (20 mM Tris, pH 7.5, 10 mM KCl, 1 mM DTT, 20 mM PMSF and 0.2 mM EDTA), incubated at 4 °C for 24 h, and dialysing for 4 h against 20 mM Tris, pH 7.9, plus 1 mM DTT. Samples were normalized to contain 0.1 mg ml⁻¹ protein, aliquoted, and stored at -80 °C.

Immunodepletion and reconstitution. Protein A-Sepharose CL-4B beads (0.5 ml; Pharmacia) were washed with PBS and mixed with an equal volume of anti-Urp serum or preimmune serum. After binding overnight, antibodies were crosslinked to protein A according to ref. 22. Antibody-bound beads were washed with 10 ml 0.1 M glycine (pH 2.5) then 2 × 10 ml buffer D (20 mM HEPES pH 7.6, 100 mM KCl, 1 mM DTT, 10% glycerol). HeLa cell nuclear extracts were used for immunodepletion as previously described³. To minimize co-depletion of other essential splicing factors, high-salt conditions were used: 1 ml nuclear extract was adjusted to 0.5 M KCl, then passed through a 0.5-ml (packed volume) antibody-affinity column 5 times at 4 °C. The depletion procedure was repeated once on a fresh 0.5-ml affinity column. Depleted extracts were dialysed against buffer D for 4 h at 4 °C, aliquoted, and stored at -80 °C. Antibody-affinity columns were regenerated by washing with 0.1 M glycine (pH 2.5) and stored in buffer D. Standard splicing reactions were carried out as described³.

In vitro and in vivo binding. U2AF⁶⁵ and U2AF³⁵ cDNAs in pSP64 and in pGEX were provided by J. Fleckner and M. Green; anti-U2AF⁶⁵ and U2AF³⁵ antisera were gifts from P. Zuo and T. Maniatis, and were used at 1:5,000 dilution. Wild-type and mutant Urp cDNAs generated by PCR were inserted downstream of the T7 promoter in pcDNA3 (Invitrogen). *In vitro*-translated proteins were prepared using a TNT kit (Promega) in the presence of [³⁵S]methionine. ³⁵S-labelled proteins were incubated with 1 µg GST-U2AF⁶⁵ or GST-U2AF³⁵ immobilized on 5 µl glutathione beads (Pharmacia) in binding buffer (20 mM HEPES, pH 7.6, 3 mM MgCl₂, 0.1 mM EDTA, 0.1% Tween, 10% glycerol, 1 mM DTT, 0.4 M KCl) plus 5% BSA at 4 °C for 1 h. After beads were washed 4 times in binding buffer, proteins were resolved on SDS-PAGE, visualized by autoradiography, and quantified on a Phosphorimager.

Two-hybrid pairwise interactions were carried out using the LexA system²³ in the host yeast strain EGY48. Bait plasmids contained individual cDNAs in frame downstream of the LexA gene in pEG202, and prey plasmids had individual cDNAs in pJG4-5. Several bait and prey plasmids (U2AF³⁵ and SF2/ASF) were gifts from J. Wu³. EGY48 cells were co-transformed by the lithium acetate method and transformants were selected on glucose/CM-ura-his-trp plates. Six randomly picked colonies were restreaked on Gal/Raff/CM-ura-his-trp plates containing X-gal for colour development. Interactions were quantified by liquid β-galactosidase assay²³.

Received 14 April; accepted 21 May 1997.

1. Ruskin, B., Zamore, P. D. & Green, M. R. A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell* **52**, 207–219 (1988).
2. Krämer, A. The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annu. Rev. Biochem.* **65**, 367–490 (1996).
3. Wu, J. & Maniatis, T. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75**, 1061–1070 (1993).
4. Kohetz, J. D. et al. Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature* **368**, 119–124 (1994).
5. Zuo, P. & Maniatis, T. The splicing factor U2AF³⁵ mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes Dev.* **10**, 1356–1368 (1996).
6. Roscigno, R. F. & Garcia-Blanco, M. A. SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome. *RNA* **1**, 692–706 (1995).
7. Fetzter, S., Lauber, J., Will, C. L. & Lührmann, R. The [U4/U6.U5] tri-snRNP-specific 27K protein is a novel SR protein which can be phosphorylated by the snRNP-associated protein kinase. *RNA* **3**, 344–355 (1997).
8. Gui, J.-F., Lane, W. S. & Fu, X.-D. A serine kinase regulates intracellular localization of splicing factors in the cell cycle. *Nature* **369**, 678–682 (1994).
9. Gui, J.-F., Tronchère, H., Chandler, S. D. & Fu, X.-D. Purification and characterization of a kinase specific for the serine- and arginine-rich pre-mRNA splicing factors. *Proc. Natl Acad. Sci. USA* **91**, 10824–10828 (1994).
10. Colwill, K. et al. SRPK1 and Clk/Sty protein kinases show distinct substrate specificities for serine/arginine-rich splicing factors. *J. Biol. Chem.* **271**, 24569–24575 (1996).
11. Fu, X.-D. The superfamily of arginine/serine-rich splicing factors. *RNA* **1**, 663–680 (1995).
12. Manley, J. L. & Tacke, R. SR proteins and splicing control. *Genes Dev.* **10**, 1569–1579 (1996).
13. Kitagawa, K. et al. Isolation and mapping of human homologues of an imprinted mouse gene U2af1-rs1. *Genomics* **30**, 257–263 (1995).
14. Nabetani, A., Hatada, I., Morisaki, H., Oshimura, M. & Mukai, T. Mouse U2af1-rs1 is a neomorphic imprinted gene. *Mol. Cell. Biol.* **17**, 789–798 (1997).
15. Hyashizaki, Y. et al. Identification of an imprinted U2af binding protein related sequence on mouse chromosome 11 using the RLGs method. *Nature Genet.* **6**, 33–40 (1994).
16. Yamaoka, T., Hatada, I., Kitagawa, K., Wang, X. & Mukai, T. Cloning and mapping of the U2af1-rs2 gene with a high transmission distortion in interspecific backcross progeny. *Genomics* **27**, 337–340 (1995).
17. Zamore, P. D. & Green, M. R. Biochemical characterization of U2 snRNP auxiliary factor: an essential pre-mRNA splicing factor with novel intranuclear distribution. *EMBO J.* **10**, 207–214 (1991).
18. Rudner, D. Z., Kanaar, R., Breger, K. S. & Rio, D. C. Mutations in the small subunit of the Drosophila U2AF splicing factor cause lethality and developmental defects. *Proc. Natl Acad. Sci. USA* **93**, 10333–10337 (1996).
19. Zamore, P. D. & Green, M. R. Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. *Proc. Natl Acad. Sci. USA* **86**, 9243–9247 (1989).
20. Zhang, M., Zamore, P. D., Carmo-Fonseca, M., Lamond, A. I. & Green, M. R. Cloning and intracellular localization of the U2 small nuclear ribonucleoprotein auxiliary factor small subunit. *Proc. Natl Acad. Sci. USA* **89**, 8769–8773 (1992).
21. Hannon, G. J., Dmetrick, D. & Beach, D. Isolation of the Rb-related p130 through its interaction with CDK2 and cyclins. *Genes Dev.* **7**, 2378–2391 (1993).
22. Harlow, E. & Lane, D. *Antibodies: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, New York, 1988).
23. Zervos, A. S., Gyuris, J. & Brent, R. Mxi, a protein that specifically interacts with Max to bind Myc-Max recognition sites. *Cell* **72**, 223–232 (1993).

Acknowledgements. We thank G. Hannon for yeast two-hybrid libraries; J.-L. Guan for raising anti-Urp antibodies; J. Fleckner and M. Green for U2AF cDNAs; P. Zuo and T. Maniatis for anti-U2AF antibodies and purified U2AF³⁵ protein; J. Wu for two-hybrid plasmids and for advice; and members of our laboratory, especially H.-Y. Wang and J. Dyck, for suggestions and contributions. X.-D.F. is a Searle Scholar. H.T. was supported by postdoctoral fellowships from the Human Frontier Science Program and La Société de Secours des Amis de la Science. J.W. was supported by a postdoctoral fellowship from the NIH. This work was supported by grants to X.-D.F. from the NIH.

Correspondence and requests for materials should be addressed to X.-D.F. (e-mail: xdfu@ucsd.edu).

correction

Expression of *Radical fringe* in limb-bud ectoderm regulates apical ectodermal ridge formation

Ed Laufer, Randall Dahn, Olivia E. Orozco, Chang-Yeol Yeo, Jacqueline Pisenti, Domingos Henrique, Ursula K. Abbott, John F. Fallon & Cliff Tabin

Nature **386**, 366–373 (1997)

We inadvertently omitted to acknowledge D. Ish-Horowitz for the Ser-2 clone and for supporting D.H. (through funds awarded by the ICRF and HFSP), and T. Vogt, S. H. Johnston, C. Rauskolb, E. Wieschaus, K. Irvine, S. Egan, Y. Rao, Y. Wang and P. Leder for sharing unpublished results. □

Using GFP to study virus invasion and spread in plant tissues

GFP is beginning to revolutionize the study of virus movement in plants. Insertion of the *gfp* gene into the viral genome allows the virus to be tracked both in whole plants and also in single cells.

Karl J. Oparka, Alison G. Roberts, Simon Santa Cruz, Petra Boevink, Denton A.M. Prior and Anna Smallcombe

Using non-destructive imaging techniques, it has proved possible to devise methods for studying the spread of plant viruses at both the cellular and whole-plant level using a combination of low-magnification lenses and confocal laser scanning microscopy (CLSM). This technique has made it possible to gain valuable insights into the mechanisms of virus spread through plant tissues and to study the role of specific viral genes in this process.

In the past, attempts to unravel the role of virus-encoded genes have employed the marker protein β -glucuronidase (GUS)¹, which enabled histochemical staining to localize virus in specific tissues and cells. However, this procedure is inherently destructive and research has been awaiting a non-invasive *in vivo* marker of gene expression. The green fluorescent protein (GFP) from *Aequorea victoria* has the advantage that it is strongly fluorescent and requires no substrates or cofactors, other than molecular oxygen, to form the fluorescent molecule².

The *gfp* gene can be incorporated into viral genomes for expression as a free protein to act as a reporter of viral infection³. Alternatively, GFP can be fused to selected proteins in order to reveal their cellular and subcellular localizations^{4–10}. GFP appears to be non-toxic to plants, is highly stable inside cells and photobleaches very slowly, making it suitable for long-term *in vivo* fluorescence imaging.

We have utilized GFP in a variety of ways to investigate the viral infection process with the aim of devising methods that will eventually prevent or limit the viral invasion of crop plants. This investigation has been greatly



Figure 1 A *Nicotiana benthamiana* plant systemically infected with potato virus X (PVX) expressing the green fluorescent protein (GFP).

assisted by the use of the confocal microscope, and emphasis is placed on the advantages conferred by using low-power, as well as the more usual high-power objective lenses.

The movement of viruses from cell to cell occurs via plasmodesmata, pores that connect one plant cell with another. This is an active process requiring virus-encoded proteins, termed movement proteins (MPs) and, in some viruses, also requires the coat protein (CP). Plasmodesmata are too small to allow the passage of either virus particles or viral nucleic acid, but one role of MPs is to activate a 'gating' mechanism, opening the plasmodesmata and allowing cell-to-cell spread of infection¹⁰.

In addition to a role in plasmodesmatal modification, a number of other functions have been proposed for viral MPs, including nucleic acid binding¹¹, association with the cytoskeleton¹² and trafficking viral genomes to and through the plasmodesmata. However, it should be noted that different viruses utilize a number of distinct methods for cell-to-cell movement^{12,13}.

This report describes the use of a potato virus X-based vector to express the GFP either as a free cytosolic protein or as a fusion to both homologous and heterologous viral proteins. GFP can be detected at both macroscopic and microscopic levels but here the emphasis will be made on the microscopic techniques involved in the study of virus invasion and spread and the subcellular localization of GFP-tagged proteins.

GFP constructs

A number of viral constructs and PVX-based expression vectors have been prepared (described in refs 2 and 4): in all cases the original GFP gene described by Chalfie *et al.*¹⁴ was employed. (1) PVX expressing free GFP to enable viral spread to be monitored; (2) PVX expressing a GFP-coat protein fusion that assembles into fluorescently labelled virus particles, enabling their subcellular localization to be determined; (3) PVX expressing free GFP but with a coat protein deletion to determine the role of the CP in viral movement; (4) PVX vector expressing a fusion of GFP to the MPs of tobacco mosaic virus (TMV) and cucumber mosaic virus (CMV) to determine the localization of these MPs during viral spread; (5) PVX vector expressing GFP targeted to the endoplasmic reticulum (ER) to allow the ER network to be imaged.

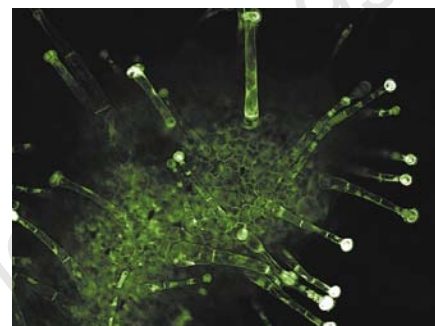


Figure 2 Leaf trichome cells under the confocal laser scanning microscope (CLSM). The strong GFP signal from individual cells represents the presence of replicating virus.

Infection and microscopic examination

Plants were infected with PVX containing various GFP inserts using a standard manual virus inoculation technique. Infected plant sap, purified virus particles, transcribed RNA or plasmid DNA can all be used as inoculum to initiate a viral infection. At a macroscopic level, the progress of the infection by cell-cell spread (local movement) and systemically (long distance movement) was followed by examining whole plants under long-wavelength ultraviolet illumination (365 nm) supplied by a 'Blak Ray' hand-held lamp (Ultraviolet Products, Cambridge, UK). Whole-plant photography was carried out with a 35-mm camera equipped with a green $\times 1$ filter (Hoya, Japan) in order to remove the autofluorescence associated with chlorophyll (Fig. 1). It should be noted that if wavelength-shifted GFP mutants are used, no fluorescence will be detected under UV light. Leaves were then selected for microscopic examination using a Bio-Rad

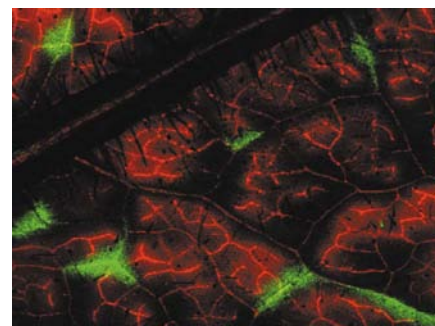


Figure 3 A dual-labelled leaf in which Texas Red dextran was taken up into the veins through the cut petiole (red). The GFP-labelled PVX (green) is seen to be unloading from the veins and spreading into the surrounding mesophyll tissues.

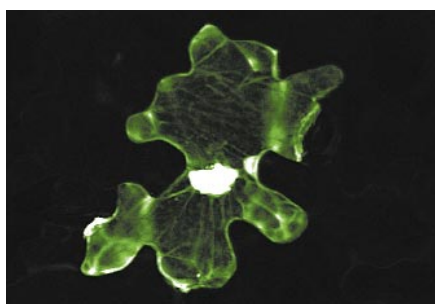


Figure 4 A single infected epidermal cell of *N. benthamiana* containing the PVX vector but lacking the coat protein (CP). The virus is unable to spread from the inoculated cell in the absence of the CP.

MRC-1000 confocal laser scanning microscope equipped with a krypton/argon laser (Bio-Rad, Hemel Hempstead, UK).

In some cases, the petioles of excised leaves were placed in a solution of 3 K Texas Red dye (Molecular Probes, Eugene, Oregon, USA). The dye was carried in the xylem through the transpiration stream and clearly labelled the vascular system of the leaves¹⁵. Texas Red and GFP can be imaged simultaneously with the confocal microscope, using 568-nm and 488-nm excitation wavelengths, respectively. For simultaneous imaging, both excitation lines were used and filter blocks T1 and T2A were positioned in the scan head. Whole leaves were supported on the x-y stage of the microscope (standard Nikon Optiphot attached to the MRC-1000), and the cut end of the petiole was placed in a water reservoir to allow transpiration to continue during imaging.

Localization of GFP

Leaves that were inoculated with PVX expressing GFP as a free cytosolic protein developed isolated infection foci, which initiated from single cells and expanded radially as the virus moved from cell to cell. The infected areas containing GFP were highly fluorescent and easily visible when viewed under long-wavelength UV light. In the case of whole plants, GFP infection sites on both inoculated and systemically infected leaves were clearly visible with the naked eye (Fig. 1). Under the confocal microscope individual cells or groups of cells, such as the

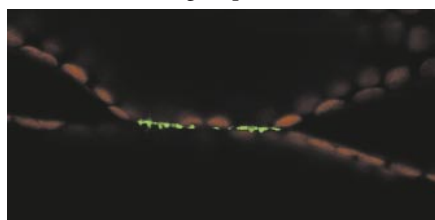


Figure 5 Localization of the viral movement protein (MP) of cucumber mosaic virus (CMV) expressed as a fusion to GFP. The MP-GFP fusion has been targeted to plasmodesmata in the walls between adjoining leaf mesophyll cells. Note the red autofluorescence from chloroplasts.

leaf trichomes shown in Fig. 2, were seen clearly to contain GFP-expressing virus. Continued cell-cell movement of virus on the inoculated leaf brings it into contact with vascular tissues, allowing the virus to enter the phloem and be transported around the plant with the flow of photoassimilate. The subsequent exit of the virus from the phloem in systemically infected tissues was then imaged using the confocal microscope. Infected leaves were additionally labelled with Texas Red dye allowing the sites of virus exit and subsequent spread to be correlated with the underlying vein network¹⁵ (Fig. 3).

When GFP was fused to the CP gene of PVX, the assembled virus possessed a GFP 'overcoat'⁴. The GFP then attached to the viral particles and was not produced free in the cytoplasm. Microscopic imaging of cells infected with the overcoated virus showed that the virus particles aggregated to form inclusions that could easily be detected using the confocal technique⁴. Alternatively, deletion of the CP gene, replacing it with GFP, resulted in the virus being restricted to single, inoculated cells (Fig. 4). This demonstrated that the CP is essential for the cell-to-cell movement of PVX.

Leaves inoculated with a PVX vector that contained a fusion of GFP to the TMV MP showed that GFP became localized to punctuate spots on the cell wall shown to be plasmodesmata⁶. Similarly, the MP of CMV has been expressed from the PVX vector⁵ and shown to be localized to plasmodesmata. In Fig. 5, a cluster of plasmodesmata labelled with MP-GFP fusion protein are depicted between two adjacent leaf palisade mesophyll cells. While it is clear that MPs are a prerequisite for cell-cell spread of viruses, and that one of the functions of MPs may be to widen the plasmodesmal pore, the mechanism has not been fully elucidated.

In addition to fusing GFP to viral proteins in order to determine their role in the infection process, non-viral proteins can also be expressed from the PVX vector in plant cells. In cell-biological studies, virally expressed GFP was delivered successfully to the ER using the appropriate ER targeting and retention signals¹⁶ (Fig. 6).

Advantages of using GFP

The ability to trace non-invasively the spread of the virus in intact plants is a technique that has the potential to advance greatly the study of viral infection in plants. Since the discovery of GFP as a molecular marker, researchers have a new tool that allows long-term *in vivo* analysis of the progression of viral infections. It is also now being used to identify the roles of different viral-encoded proteins in the subcellular events associated with the infection process.

For the type of work presented here, it is important to be able to image a large field of view under low magnification. Both GFP and

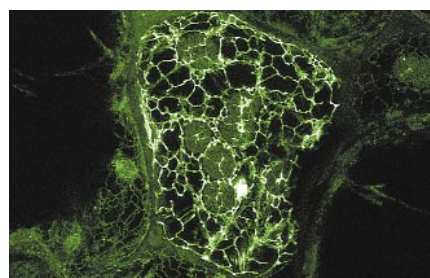


Figure 6 Virus-mediated delivery of GFP to the endoplasmic reticulum of leaf mesophyll cells. The PVX vector was used to express GFP containing the appropriate ER targeting and retention signals.

Texas Red have been found to be sufficiently fluorescent to allow the use of low-power lenses. For example, the photograph presented in Fig. 3 was imaged under a $\times 2$ lens, and even $\times 1$ lenses can be used effectively in conjunction with confocal microscopy. At such low magnifications, the confocality is obviously limited, but nonetheless the system retains the advantages of digitized imaging while removing out-of-focus flare to produce sharp images from which valuable information can be obtained. □

K.J.O., A.G.R., S.S.C., P.B. & D.A.M.P. are at the Unit of Cell Biology, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK. A. S. is at Bio-Rad Microscopy Division, Maylands Avenue, Hemel Hempstead HP2 7TD, UK.

1. Dolja, V.V. *et al.* *Proc. Natl Acad. Sci. USA* 89, 10208–10212 (1992).
2. Prasher, D.C. *TIG* 11, 320–323 (1995).
3. Baulcombe, D.C. *et al.* *Plant J.* 7, 1045–1053 (1995).
4. Santa Cruz, S. *et al.* *Proc. Natl Acad. Sci. USA*, 93, 6286–6290 (1996).
5. Oparka, K.J. *et al.* *Trends Plant Sci.* 1, 412–418 (1996).
6. Oparka, K.J. *et al.* *Protoplasma* 189, 133–141 (1995).
7. Epel, B.L. *et al.* *Gene* 173, 75–79 (1996).
8. Padgett, H.S. *et al.* *Plant J.* 10, 1079–1088 (1996).
9. Heinlein, M. *et al.* *Science* 270, 1983–1985 (1995).
10. Oparka, K.J. *et al.* *Plant J.* (in the press).
11. Citovsky, V. *et al.* *The Plant Cell* 4, 397–411 (1992).
12. Maule, A.J. *Crit. Rev. Plant Sci.* 9, 457–473 (1991).
13. Deom, C.M. *et al.* *Cell* 69, 221–224 (1992).
14. Chalfie, M. *et al.* *Science* 263, 802–805 (1994).
15. Roberts, A.G. *et al.* *The Plant Cell* (in the press).
16. Boevink, P. *et al.* *Plant J.* 10, 935–941 (1996).

Acknowledgements. This research is funded by The Scottish Office Agriculture, Environment and Fisheries Department.

ADVERTISEMENT

2-Hybrid System Media

Over 400 recipes
and formulae in stock

BIO 101 has the largest selection
of Bacterial & Yeast Media available in
capsule, powder & pouch form

Call us or your local distributor for ordering & information:
(800) 424-6101 ♦ fax: (760) 598-0116
e-mail: info@bio101.com ♦ www.bio101.com

CONVENIENCE

CONSISTENCY

ECONOMY

RELIABILITY

SELECTION

With the objective of reviewing recent releases in the field of microscopy, this feature looks at a variety of optical and confocal microscopes, instruments for SPM, SEM, AMF and TEM, as well as accessories and software.

PIM-III inverted microscope

From World Precision Instruments

A microscope designed for a wide range of laboratories and biological applications

WPI's new microscope comes with high-quality, long-working distance plan achromatic objectives. The optics are manufactured to provide a high-resolution, flat viewing field and sharp imaging. A large, easily accessible stage permits external manipulators and other apparatus to be located close to the specimen. The PIM-III comes complete with attachments for photomicrography, including a photo tube with built-in $\times 5$ optics, a T-mount adapter (for 35-mm SLR cameras) and a $\times 10$ viewfinder eyepiece. Video output allows group viewing of specimens, and a permanent record may be obtained through use of a VCR or computer file. Also included are yellow, green, blue and gray neutral-density filters.

Reader Enquiry No. 100

2-photo microscopy

From Bio-Rad

Laser microscopy to aid in the study of the internal mechanisms of living cells

Bio-Rad claim that this technology enables researchers not only to observe cells for longer, but also to observe deeper into tissues and organisms. Developed from laser scanning confocal microscopy, this technique is said to be an improvement as it allows the study of live cells for extended periods of time without photodamage. Additionally, the laser wavelength used for 2-photon work is stated to penetrate deeper into tissue (200–300 μm), thereby extending the possible range of experiments.

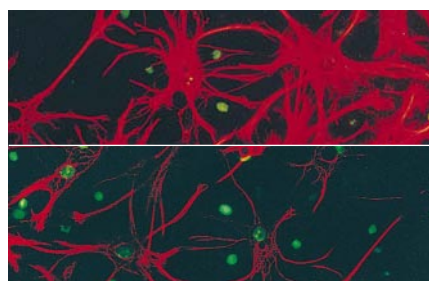
Reader Enquiry No. 101

AX REXBA accessory

From Olympus

A device that is designed to deal with the varying intensities of individual fluorochromes in multiple fluorescence excitation imaging

This accessory has been designed for the AX70 Provis research microscope, giving better results with multiple excitation, which enables specific, individual colours, such as FITC plus TRITC, PI or Texas Red, to be attenuated or intensified. It is often impossible to ensure when preparing the specimen that the intensity of the individ-



These pictures contrast multiple fluorescence with (bottom) and without (top) AX REXBA.

ual colours is at the same level. By simply turning the knob of the new device, the brightness of the dominating colour is continuously attenuated until the relative intensity of the colours meets the user's requirements. The two photographs of a culture of rat brain cells show the effect, says Olympus. These are stained with FITC and TRITC. The top photo shows the image without the use of the device. The FITC signal is dominated by the TRITC signal. The accessory was used for the bottom photo, helping to attenuate the TRITC signal far enough to make the FITC signal more clearly visible. It is even possible to suppress one of the two signals completely if only one colour is to be recorded.

Reader Enquiry No. 102

AMF/SEM/SPM/TEM

BioScope AFM

From Digital Instruments

An AFM designed specifically for biologists

This atomic force microscope (AFM) was created from a modified, inverted optical microscope by adding an AFM stage, which provides high rigidity and low vibration, thereby reducing noise and improving image quality. It offers biologists a familiar platform for AFM work by integrating an AFM with popular inverted optical microscopes. The stage also provides rigid mounts for standard Petri dishes, microscope slides and cover slips, simplifying sample mounting and requiring little or no sample preparation. Because the optical microscope views the sample from underneath, manipulating samples and aligning the AFM probe over structures of interest is easier and more convenient. The AFM operates from above with concurrent optical and video presentation so that the user can observe the sample as it is imaged by the AFM. Dynamic processes can be observed, using automated repetitive imaging. The AFM and optics provide an effective magnification range from 25 to 1×10^7 times. By simply removing the lightweight AFM unit, users can have full access to the optical microscope. The AFM can also be used separate from the optical microscope, or combined with custom AFM assemblies. The manufacturer states that users can change

from operation in air to fluids in a matter of a few minutes.

Reader Enquiry No. 103

TOPSystem 3

From Oxford Instruments

Scanning probe microscopy (SPM) control system written to run under Windows95

According to the manufacturer, the low noise and high stability of this instrument is particularly suited to high-resolution spectroscopy. The system's software is said to provide easy-to-use and flexible routines for all commonly used spectroscopic functions. It offers enhanced data acquisition and image processing with an intuitive graphical interface. The modular PC-based system eliminates the need for a workstation with hardware that features tunnelling below 2 pA, and simultaneous multi-channel acquisition with flexible acquisition software. An interactive status panel allows SPM parameters to be set and adjusted, even during scanning. The system can control a range of SPM heads, including those made by users for specific applications. The company has an upgrade package for users of TOPSystem II.

Reader Enquiry No. 104

H-7500 TEM

From Hitachi

A TEM with a Windows interface to simplify biological imaging and analysis

The microscope permits high-resolution imaging and detailed analysis functions via easy mouse operation and a clear, graphical Windows environment. It allows direct, full-area observation of a large 160-mm diameter screen at only $\times 700$ magnification. Fields of interest and beam positions can be easily selected because the instrument's Hiper-Stage is linked to magnification, resulting in a constant specimen-traverse speed independent of magnification. The TEM also incorporates a half-frame photo capability that, beyond extending film capacity, permits combined survey/zoom operations and preparation of stereo-pair film records. Other features include a super-low-dose mode for minimal beam damage, an automatic pre-irradiation system to prevent specimen drift and damage to supporting membranes, an



The H-7500TEM now offers a Windows interface.

automatic shutter for automated sequential photo recording and TV capability permitting direct image acquisition for monitor display, as well as external video recording and/or processing.

Reader Enquiry No. 105

XL30 ESEM

From Philips Electron Optics

A tungsten and LaB6 gun for SEM microscopy

This instrument uses the company's ESEM technology for true secondary electron imaging at water vapour pressures of 1.3 kPa or higher. The tungsten/LaB6 guns complement the existing FEG (field-emission gun). Moreover, the ESEM mode eliminates the high vacuum in the microscope chamber of conventional SEMs, using a higher pressure gaseous atmosphere for study of wet or oily samples. An enhanced controlled pressure mode with a small beam skirt uses a backscatter detector for imaging at chamber pressures below ~200 Pa. The instruments also work in conventional high-vacuum SEM mode, with optional SE detector. The ESEM provides high-magnification SE images of water-containing samples, extending secondary imaging well into the 1.3 kPa vacuum range (where hydrated samples can be stable). A built-in water reservoir provides the water vapour, or the operator can select an external gas like nitrogen or air.

Reader Enquiry No. 106

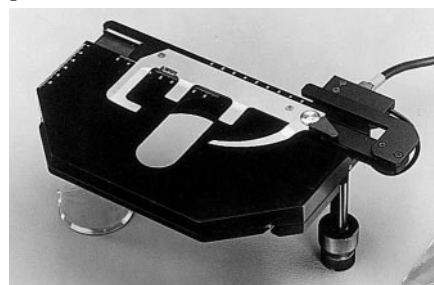
Accessories

Digitally Encoded Stage

From Carl Zeiss

This manual, mechanical stage uses a electronic Vernier scale with digital encoding

For use with the Axioplan 2 digitally controlled microscope, this new stage is coated with ceramic for maximum wear and resistance to chemical erosion. Stage movement covers a 75-mm × 50-mm area, and the Vernier scale can be displayed either on the display unit or by linkage to a PC. Coordinates can be accurately logged from the display unit, or recorded directly into the system's software. This product offers an alternative to a motorized stage for digital recording of specimen position, and complements the 'mark and find' software



Zeiss' stage features an electronic Vernier scale.

developed for the new Axioplan/Axiophot 2 microscopes. Where communication is solely with a PC, an SCPC board is used, eliminating the need for the external display unit. Alternatively, the system can be coupled to a laptop computer via a PCMCIA card.

Reader Enquiry No. 107

Auto-Montage software

From Synoptics

A motorized z-stepper option has been added to this software

The introduction of a motorized Z-stepper for driving the fine focus knob of optical microscopes allows a series of images at different focal planes to be acquired automatically, thereby providing total automation of the Auto-Montage sequence. The software helps to eliminate the problems associated with inadequate depth of focus in optical microscopy, combining the in-focus portions of a series of images of the same field of view, acquired at different focal planes. Full control allows the user to specify the top and bottom limits of the focal range and the number of steps between images. Used in conjunction with an integrated framestore board, the system acquires a sequence of source images for the processing software at the different focal planes, storing and numbering them in the process. These source images are compared by the processing software using a user-selectable range of special algorithms to give an intelligent interpretation of the in-focus areas from each image.

Reader Enquiry No. 108

KS 400 image analysis software

From Kontron Elektronik

Two- and three-dimensional image analysis software optimizes laboratory processes

Now available in North America, this software has a modular configuration that enables users to purchase a system to meet their budgets, and still have the option to add capabilities later. The KS 400 is suited to material analysis, industrial inspection, process and quality control, cell and tissue analysis, biomedical and forensics microscopy (light-optical, laser-scan, and electron microscopy), and knowledge-based classification and three-dimensional reconstruction. The system is Windows compatible and permits classification of objects according to the desired parameters, checking divergence automatically. All inputs can be entered by mouse or digitizer, which is advantageous when the user wants to measure object contours. A macro interpreter allows common command sequences to be recorded and then played back automatically, which is said to help eliminate operator key-stroke errors. The KS 400 incorporates contour-based measuring

algorithms and fast object-specific and field-specific measurements.

Reader Enquiry No. 109

CFI60 infinity optical system

From Nikon

Two new series of objectives designed for use with Eclipse microscopes

There are 12 new plan achromat and 10 new achromat flat-field objectives in Nikon's CFI60 infinity optical system. The new objectives are designed for microbiologists, histologists and other life science professionals, and are said to deliver crisp, bright, high-contrast images across a field of view that is 20 per cent larger than previous imaging systems offered. The new objectives are designed for brightfield, darkfield and fluorescence techniques, and offer the highest numerical apertures with the longest working distances yet offered by the company. The plan achromat objectives include a ×0.5 objective for macro viewing. This cone-shaped objective allows the user to image a specimen and document it on 35-mm film at a 1:1 ratio.

Reader Enquiry No. 110

These notes are compiled by Brendan Horton from information provided by the manufacturers. For more details, fill in the reader service card bound inside the journal.

ADVERTISEMENTS



PEPTIDE TECHNOLOGIES

- Custom Peptide Synthesis
- Purified and Analyzed for Your Research
- Biotinylated, Dye-labelled, Phosphopeptides, etc.
- Peptide Vaccines: KLH Conjugates and MAPS
- CALL 301 869 7306 (USA)
- FAX 301 869 7308
- URL <http://www.luxnet.com/peptide/>

Check out
our homepage at
<http://www.resgen.com>

Peptides
\$18/residue

Antibodies
monoclonal - \$8500
polyclonal - \$1195

Complete services for your custom peptide and polyclonal or monoclonal antibody needs.

Research Genetics, Inc.
U.S. or Canada 800-533-4363
U.K. 0-800-89-1393
FAX 205-536-9016